Portal NURC-SP: Design, Development, and Speech Processing Corpora Resources to Support the Public Dissemination of Portuguese Spoken Language

Ana Carolina Rodrigues[1], Alessandra A. Macedo[2], Arnaldo Candido Jr[3], Flaviane R. F. Svartman[4], Giovana M. Craveiro[1], Marli Quadros Leite[4], Sandra M. Aluísio[1], Vinícius G. Santos[4], Vinícius M. Garcia[5]

[1]Institute of Mathematics and Computer Science, University of São Paulo
[2]Faculty of Philosophy, Sciences and Letters at Ribeirão Preto, University of São Paulo
[3]Institute of Biosciences, Letters and Exact Sciences, São Paulo State University
[4]Faculty of Philosophy, Languages and Literature, and Human Sciences, University of São Paulo
[5]Ocean Technologies

ana2.rodrigues@alumni.usp.br, ale.alaniz@usp.br, arnaldo.candido@unesp.br, flavianesvartman@usp.br, giovana.meloni.craveiro@alumni.usp.br, mqleite@usp.br, sandra@icmc.usp.br, vinicius.santos@alumni.usp.br, vinicius.molina.garcia@alumni.usp.br

Abstract

We present the Portal NURC-SP Digital, an interactive web-based repository designed to maintain, organize, and facilitate access to the NURC-SP corpora collection. One of the objectives of the work on the NURC-SP corpora was to evaluate speech processing tools that allowed rapid data processing to make all NURC-SP material available on a public and dedicated portal. This objective was only possible with the joint effort of researchers working in Prosody and Speech Processing of Brazilian Portuguese. NURC-SP Digital continues a similar project called NURC Digital, making data processing fast and available for linguistic research and training speech processing models, two objectives for which we design Portal NURC-SP. In this paper, we present the status of the data processing of NURC-SP and make the URL of the Portal publicly available to allow users to have their experience accessing this large data of audio files aligned with speaker-aware transcripts, including rich metadata.

1 Introduction

NURC-SP was the Sao Paulo division of the NURC — Cultured Linguistic Urban Norm (*Norma Urbana Linguística Culta*), a project that began in 1969 to document and study Portuguese spoken language by people with a high degree of formal education in five Brazilian capitals: Recife, Salvador, Rio de Janeiro, Sao Paulo, and Porto Alegre. Located at the University of Sao Paulo (USP-FFLCH), NURC-SP collected more than 300 hours of Sao Paulo speakers throughout the 1970s. Its collection of oral records has been extensively used in various studies of spoken language and resulted in 3 volumes containing the transcription of their shared corpus, also known as the Minimum Corpus, and a series of 14 books on different topics such as linguistic variation, relationships between text and speech and typical features of orality (Silva, 1996).

This rich audio material was stored in magnetic tapes at the time, complicating access and modern use. Many of the studies from the NURC Project derive from transcriptions of part of the recorded audio selected by researchers in each city where the project had a center (Oliveira Jr., 2016).

Advances brought by the Internet and the development in computer power and memory availability made it possible to store language collections in digital mediums, and search data tools and collaborative platforms were created to group them, such as Kaggle, HuggingFace and Google Data Search. Regarding language-driven ones, the CLARIN Virtual Language Observatory (Clarin VLO) offers access to a broad range of language data, and specifically for Portuguese, the Portulan Clarin (Branco et al., 2020) provides a repository of language resources and a workbench of tools[1]. Additionally, the availability of language data has become a necessity not only in Linguistics but also in Speech Processing studies for the development of tools, such as (i) automatic speech recognition (ASR) that automatically transcribes speech, (ii) multi-speaker

[1] https://portulanclarin.net/

synthesis text to speech (TTS) that generates several voices from different speakers, and (iii) diarization that breaks down an audio stream of multiple speakers into segments corresponding to the individual speakers. Following the advances, from 2014 to 2017, NURC-SP had its original analog audios digitized by the Alexandre Eulalio Center for Cultural Documentation (CEDAE/UNICAMP) and in December 2020 made available to the TaRSila Project as a base to build training datasets for spontaneous speech recognition systems and facilitate future language studies, through the availability of a portal with specific searching tools.

As a result, TaRSila Project started NURC-SP Digital, a joint multidisciplinary work to improve, share, and develop new material for NURC-SP. Three subcorpora are being produced within the initiative, and the development of a dedicated portal to hold them along NURC-SP collection and memory was put into practice, the Portal NURC-SP Digital.

The three subcorpora that integrate the NURC-SP Digital repository — the Minimum Corpus (MC), the Corpus of Non-Aligned Audios and Transcriptions (CATNA), and the Audio Corpus (AC) (see details in Section 3) — are the result of the TaRSila project team to process, transcribe, and carefully revise NURC-SP original audio and transcription material.

The Portal NURC-SP Digital was planned considering the needs of multiple users, and one fundamental requirement was to provide a mechanism to search the corpora collection. Providing interactive access and searching tools for corpora collection generally are not part of the corpora development flow. Much of the effort to build collections of text and audio focuses on gathering and cleaning data, letting maintenance, organization, and user's interfaces as an optional secondary tool. In Computer Science subfields such as Machine Learning, functionalities with interfaces can be a minor requirement as most researchers work directly with scripts. Consequently, the needs are met more by data volume and API open channels than by visual filtering tools. Additionally, since many studies focus on algorithm development to improve specific objective metrics, consideration regarding the particularities of the data content is unusual. On the other hand, for researchers from other fields such as Linguistics and Sociolinguistics, as well for the general public, easy access and filtering tools can determine if they know and use the material. For instance, a linguist may need to carefully analyze each sample of data in a corpus or look for a particular characteristic of a language in use.

The Portal NURC-SP Digital[2] aims to: (a) make NURC-SP audio collection available online under a license Creative Commons, specifically CC BY-NC-ND 4.0, (b) share and give easy access to the data of the three subcorpora generated within TaRSila Project, (c) provide searching tools to facilitate user interactivity with the corpora material and support future linguistics studies, and (d) preserve the memory of NURC-SP project.

2 Related Work

NURC-SP Digital has as reference the NURC Digital project from Recife[3] (Oliveira Jr., 2016) (NDRecife)[4], which proposed a method to process, organize, and provide data from NURC project. However, NURC-SP Digital differs with respect to its corpora focus, data processing and portal architecture.

While both projects share the objective of providing transcribed data in a digital format, NURC-SP Digital was thought to support future research on speech processing tools of TaRSila Project while maintaining the corpora material.

On the one hand, NDRecife decided to bring the automatic annotation of the Parser Palavras (Bick, 2000) to enrich the manual transcription and prosodic segmentation performed on Praat (Boersma and Weenink, 2023). Also, NDRecife used the web-based system TEITOK (Janssen, 2016) to allow advanced searches, including words, lemmas, part-of-speech tags, syntactic tags, morphological tags, and secondary tags (semantic information, valence, secondary word class information). On the other hand, NURC-SP Digital focus on speech processing tasks, such as automatic prosodic segmentation and ASR. MC and CATNA were annotated with automatic prosodic segmentation methods to allow fast manual revision of terminal and non-terminal prosodic boundaries[5] by annotators. Prosodic segmentation has a direct impact on ASR and TTS tools (Chen and Hasegawa-

[2] http://tarsila.icmc.usp.br:8080/nurc

[3] https://fale.ufal.br/projeto/nurcdigital/

[4] Alias adopted to avoid confusion between NURC Digital and NURC-SP Digital.

[5] Terminal boundary marks (TB) indicate the conclusion of the utterance. Non-terminal boundary marks (NTB) break of non-conclusive sequences of the utterance.

Johnson, 2004; Lin et al., 2019; Liu et al., 2022). Moreover, AC was automatically transcribed and manually revised to allow public availability of a large corpus to develop ASR models (see details in Section 3). Besides the fact that NURC-SP Digital made use of automatic tools, all data from the three corpora were manually revised.

The Portal NURC-SP Digital and its search system were developed from scratch, considering the specific material from NURC-SP Digital. With regard to the search engine, the Portal NURC-SP Digital, allows users to filter multiple features simultaneously (e.g. year=1976 **and** theme=Home **and** age group=I) and each filter displays the list of all possible labels, so users with no familiarity with the data do not have to try entries in order to know if they are part of the possible ones.

3 Corpora of NURC-SP Digital

The NURC-SP corpus is made up of 375 inquiries of three types: formal expressions (called EF), such as lectures and conference presentations; informal conversations involving speakers with a documenter present (referred to as D2), and interviews covering diverse subjects, conducted by an interviewer with the interviewee (referred to as DID). Some of the inquiries already had transcriptions — but, until then, not aligned to the audio recordings — and the vast majority were composed of only audio files. Within TaRSila Project NURC-SP was divided into three subcorpora:

- the *Minimum Corpus* (MC) (21 recordings + transcriptions) used to evaluate automatic processing methods of the entire collection (Santos et al., 2022);
- the *Corpus of Non-Aligned Audios and Transcriptions* (CATNA) (26 recordings + transcriptions); and
- the *Audio Corpus* (AC) (328 recordings without transcription), which has been automatically transcribed by WhisperX (Bain et al., 2023) that provides fast automatic speech recognition (70x real-time with the large-v2 model of Whisper (Radford et al., 2023)[6]) and speaker-aware transcripts, using the speaker diarization tool pyannote-audio[7].

MC and CATNA subcorpora were annotated with two types of prosodic boundaries — non-terminal boundaries and terminal boundaries, based on the theory and methodology used by C-ORAL-Brasil project[8] that provided studies in spontaneous speech by using phonetic-acoustic parameters and boundaries identified perceptually by trained annotators (Teixeira et al., 2018; Teixeira and Mittman, 2018; Raso et al., 2020). First, the inquiries were processed with automated methods of segmentation (Craveiro et al., 2024), then they were revised by trained annotators, using the software tool Praat. The preprocessing was responsible for preparing the textgrid[9] files making the annotation process fast and possible to be carried out by students as revising an annotation is easier than deciding the annotation from scratch. In Figure 1 we see an excerpt from an inquiry with five layers annotated in Praat, described below:

- 2 layers (TB-, NTB-) in which the speech of each speaker (-L1, -L2) and documenter (-Doc1, -Doc2) is segmented into prosodic units and transcribed according to standards adapted from the NURC Project.
- 1 layer (LA) for transcribed and segmented speech from any random speaker.
- 1 layer for comments (COM) about the audio and annotation.
- 1 layer containing the normalized version (-NORMAL) of the transcription of all TB and LA layers.
- 1 layer containing the punctuation (-PONTO) that ends each TB.

The headers of the 328 audios from AC were removed and saved for automatic metadata generation, as information about the recording of each inquiry is provided at the beginning of the audio[10]. Metadata in json format was generated with the help of ChatGPT[11] and the content of the lectures/presentations, conversations or interviews were processed by WhisperX. After that, audio and automated transcriptions were uploaded to a web-

[6] Whisper (https://openai.com/research/whisper) is an ASR trained on 680,000 hours of multilingual data collected from the web.

[7] https://github.com/pyannote/pyannote-audio

[8] www.c-oral-brasil.org/

[9] Textgrid is one of the types of objects used in Praat tool for annotation of segmentation and labelling. The resulting files from the textgrid editor in Praat have the extension ".textgrid".

[10] The information on each header is composed of: Project Name, Reel Number, Quality of Speech, Interview Topic, Number, gender and age of Informants, Names of Documenters, Date and duration of the recording, Brand of recorder and Recording conditions.

[11] https://chat.openai.com/

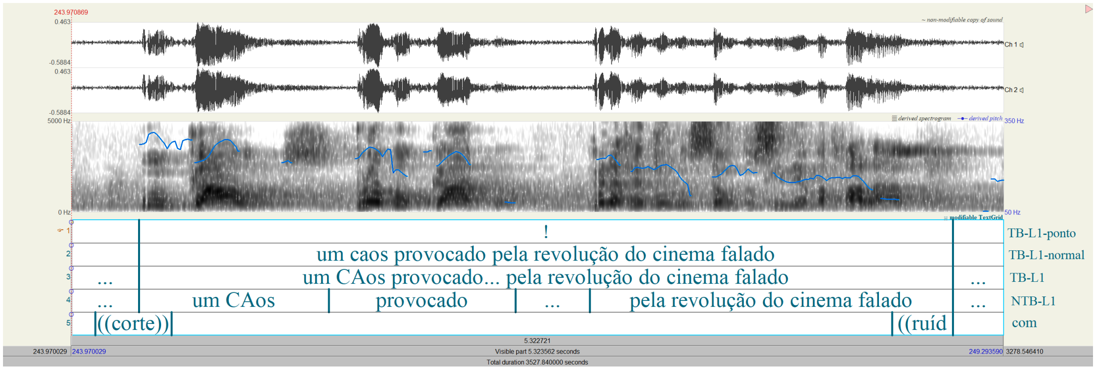


Figure 1: Excerpt from SP_EF_153 with five layers annotated in Praat.

based platform for transcription revision. The revision of automated transcriptions were performed from June 2023 to December 2023. In total, 14 annotators have worked in AC. The revision process was based on an annotation guideline designed to help making the revision uniform and contains 11 rules:

1. Orality marks were preserved in 4 cases: “né", “num", “numa”, and “tá". Other orality marks (“tá” and “tô” as a verb) and contractions (“pro”, “pra”, “cê”, among others) were transcribed following the orthographic rules.
2. Filled pauses were transcribed as close as possible to what was heard (“ah”, “ãh”, “uhum”, “aham”, among others);
3. Repetitive hesitations were transcribed (“eu fui no no mer mercado”);
4. Numbers were transcribed in words, including measurements, dates and times;
5. Individual letters were transcribed as pronounced;
6. Acronyms were transcribed as close as possible to what was spoken. For example, “i bê gê é” for “IBGE” and “USP” for “USP”. Abbreviations were expanded, according to the speaker’s pronunciation (e.g.: “kilometer” for “km”). Additionally, acronyms in English were transcribed according to the official pronunciation of the letters of the Latin alphabet in English (e.g.: “êm ái tíi” for “MIT”);
7. Foreign terms were transcribed as spelled (e.g.: laptops, netbook, notebook, among others);
8. Punctuation and capitalization were generated by Whisper, but annotators were instructed to ignore them, keeping them as received;
9. Paralinguistic sounds were noted in parentheses: (laughter), (cough), (laugh), (hiccups), (crying), among others;
10. Misunderstanding of words or passages were marked as “( )”;
11. Words truncated at the end or the beginning of the audio due to automatic segmentation failure were partially transcribed with “>” (remaining of the word is in the next audio) and “<” (start of the word is in the last audio). For example, “João” may be transcribed as “Jo>” and “<ão’ when the segmentation incorrectly breaks this word apart.

4 Design of Portal NURC-SP Digital

We have built the NURC-SP Digital Portal on the basis of the design thinking methodology (Rowe, 1991), which consists of understanding, exploring, and materializing software in diverse iterations with different versions. During the portal development, we made available a simple functional version to the NURC-SP Digital team and kept an interactive ongoing growth, using a participatory approach (Schuler and Namioka, 1993; Muller and Kuhn, 1993).

The requirements set for the portal architecture were: (i) search tools to facilitate user interactivity with the inquires of the corpora collection, (ii) easy download of the corpora files, (iii) a good page response to the user, (iv) the possibility of uncomplicated addition of new features for future improvements and (v) a friendly and intuitive interface that allows users from multiple backgrounds to access the portal. In terms of visual identity, there were no predefined requisites, therefore we began with a proposal based on the colors of the TaRSila project’s logo and ask the team members to suggest changes during the process (color, logos, section names).

In the next subsections, we present the webpage architecture, how we have addressed those requirements considering storage and performance, and the decisions we made during the NURC-SP Digital Portal development.

4.1 Architecture Overview

Figure 2 presents the architecture of the system. We used MVT architecture composed of Model, View, and Template components, implemented with Django Framework. Additionally, we made separate components to handle the background computation (Utilities and Filter properties), and data preprocessing (Data Preparation). These components work together to handle the Logic, Data and Presentation of our web portal.

In terms of the Model, some aspects needed to be handled separately. For instance, the word search was unaddressed by the relational database, and we created an internal script called directly by the View component. Also, the process of checking and preparing the metadata were performed by a separate module and data is inserted directly into the database.

Another important point for the architecture of the Portal NURC-SP Digital was to have a simple maintenance and update routine. The page needed to allow fast changes while also keeping the page available online. For security reasons, production is located on the server with a non-administrative setup environment to prevent the installation of malicious programs in case of attack by intruders, while new functionalities were tested in a local environment before being deployed. This process also served to test the flow for future improvements.

4.2 Corpora Material and Search Facilities

One crucial objective of the Portal NURC-SP Digital is to provide easy access to NURC-SP files and search tools for their features. The package for each inquiry of MC and CATNA comprises five files to be displayed and downloaded by the user:

(a) The audio recording in .wav format sampled at 48kHz, regarding the original audio digitization;
(b) The compressed version of the audio in the .mp3 format;
(c) The text transcription in .pdf. The PDF version is the original transcription made by transcribers in the 70s and 80s;
(d) The text transcription in .txt. This version is the revision of original transcriptions, made with the support of the speech analysis software Praat. Praat allows acoustic analysis of the audio, showing the oscillogram and spectrogram to support the revision carried out by annotators from the TaRSila project when performing the prosodic annotation of terminal and non-terminal segments;
(e) The text-to-speech alignment file (.textgrid format).

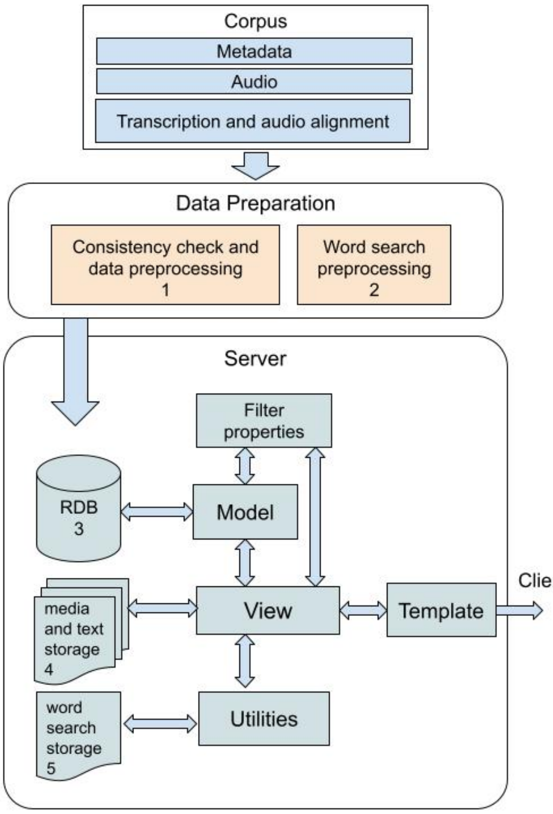


Figure 2: Portal Architecture and Data Preparation. Numbers 1 and 2 are part of the step performed before feeding data into the storage system of the portal, described in subsection 4.5 and subsection 4.6 respectively. Numbers 3, 4 and 5 are the storage described in subsection 4.4.

As for AC, the package for each inquiry comprises four files to be displayed and downloaded by the user:

(a) The audio recording in .wav format, sampled at 48kHz, regarding the original audio digitization;
(b) The compressed version of the audio in the .mp3 format;
(c) The text transcription in .txt format. This is the sequential version of the automatic transcription of WhisperX revised by annotators

in the web-based platform. It is a speaker-aware transcript, with the generic speaker generated by pyannote-audio (speaker 0, 1, 2, ...), in front;

(d) The text-to-speech alignment file (.TextGrid format).

Additionally, corpus material has a metadata file with inquiries recording conditions and speakers' information.

We provide on the corpus page of the portal the following search tools for the users:

- **Corpus filter:** The list of inquiries can be filtered using seven different characteristics: type, recording year, theme, gender, age group, audio quality and duration. They were constructed based on the metadata labels (Figure 3).
- **Corpus search by word:** Users can select inquiries by the presence of a specific token in the transcription. The result is a list of inquiries having at least one occurrence of the token in the text (Figure 3).
- **Easy download:** Users can click to download the files from the inquiries. They can also filter and make word searches to select inquiries with specific characteristics (Figure 4).

4.3 Page Response Optimization

Making decisions about where operations should take place (server-side or client-side) is an important part of web application development. They impact directly the final user perception through rendering load and time of response. On the one hand, user experience suggests computation to be executed in the front end, preventing server overload and slow feedback due to Internet traffic. On the other hand, taking security measures into account, operations cannot rely on data being checked or computed in the client system.

In the Portal NURC-SP Digital, to balance a good user experience and secure Corpora data, different solutions were implemented for each demand. All operations related to corpus data such as filtering and text searches are performed on the server-side. Inquires metadata was adapted in the front end to pre-labeled filters, in which the user chooses from a specific set of options (an example is depicted at the right side of Figure 4). This solution excludes the need for data validation when a post request is sent to the server. Inquire files for download were made available through clickable links hidden in the page (Figure 4: left). They appear by a front-end function called by users' interaction, so to avoid new page rendering while maintaining a clean visual. Additionally, the download option is performed by calling a function exclusively designed for file transferring on the server, without any need for further rendering. Other specific web features such as text reveling and pop-up windows were also selected to be executed in the client to keep the webpage fast and dynamic.

4.4 Data Storage

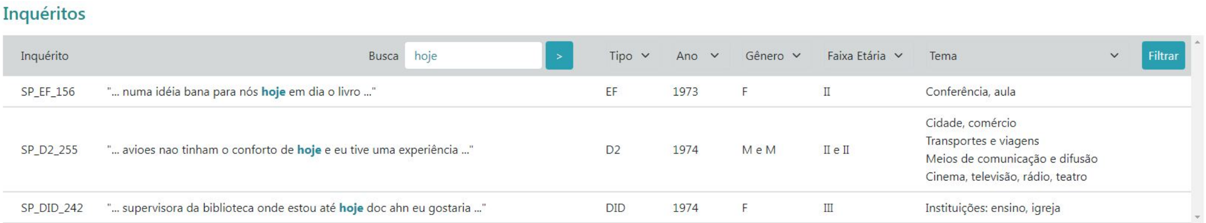
The Portal NURC-SP Digital storage system uses two distinct ways to store data in the server, based on data characteristics and corpora searches. Structured data with clear format and relationships were stored in a relational database (RDB). That is the case for each corpus metadata that feeds the filtering engine. Media and text files were stored as files with specific paths in the relational database. The corpus word search is the result of a specific strategy (described in subsection 4.6) and has data stored in json format.

The relational database tables were designed considering three functionalities: (i) accommodation of corpus metadata, (ii) optimization of queries for filtering and search requests, and (iii) data validation. Specific to the last one, database fields were implemented rigorously, fields accept only entries from a predefined labels list. Thus, each field was constrained to a set of fixed options according to the possible values of the feature (e.g., field 'type' takes only one of three entries: EF, D2, or DID).

Although this poses an extra step in which metadata must be adapted to fields' classes before database feeding, and new labels must be created in the database before being available for new entries, this design was preferable to guarantee data consistency. Moreover, it helps queries for filtering features, creating an easy relation between back-end storage and front-end presentation of data. The verification and preparation of data before putting them into the database also was revealed to be useful for tracking annotation inconsistencies and missing values.

4.5 Consistency Check and Data Preprocessing

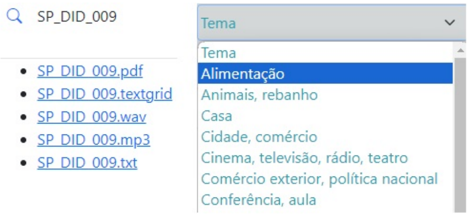
The original metadata of all subcorpora and audio transcripts of MC and CATNA were mainly obtained by manual annotation during the NURC-



Inquéritos

Inquérito	Busca hoje	Tipo	Ano	Gênero	Faixa Etária	Tema
SP_EF_156	"... numa idéia bana para nós **hoje** em dia o livro ..."	EF	1973	F	II	Conferência, aula
SP_D2_255	"... avioes nao tinham o conforto de **hoje** e eu tive uma experiência ..."	D2	1974	M e M	II e II	Cidade, comércio Transportes e viagens Meios de comunicação e difusão Cinema, televisão, rádio, teatro
SP_DID_242	"... supervisora da biblioteca onde estou até **hoje** doc ahn eu gostaria ..."	DID	1974	F	III	Instituições: ensino, igreja

Figure 3: Corpus metadata filter and search by word. Each line represents a unit from the corpus (an inquiry) and its metafeatures. From left to right: inquiry id, excerpt of the inquiry transcription with the searched word, type, year, genre of speakers, age range of speakers, divided in three groups: I (25–35), II (36–55) and III (56 or more), and theme.



SP_DID_009

- SP_DID_009.pdf
- SP_DID_009.textgrid
- SP_DID_009.wav
- SP_DID_009.mp3
- SP_DID_009.txt

Tema

Tema
Alimentação
Animais, rebanho
Casa
Cidade, comércio
Cinema, televisão, rádio, teatro
Comércio exterior, política nacional
Conferência, aula

Figure 4: Details of corpus search data facilities. Left: List of files to download. It appears with a click on the inquiry name. Right: Top part of the theme drop-down filter.

SP project gathering period. Because of the challenges inherent to the available technology of the period, and the natural variations of human annotation, some features do not follow a strict pattern and there are label discrepancies among them. Most data is being manually revised within TaRSila project, in a joint effort to bring a reliable Corpora to the public. The portal performs the last stage of metadata checks. Due to its database consistency constraints, an extra step to verify metadata was required: a consistency check and data preprocessing routine was prepared to assert metadata before putting them into the database.

A verification on AC metadata revealed that: (a) There are small differences in written annotation, for instance, the same intended theme can be slightly different as in 'ciclo *de* vida' vs. 'ciclo *da* vida'; (b) Theme labels with the same theme id vary in their text, as in 'Instituições: igreja' vs. 'Instituições: ensino, igreja', 'Diversões, esportes' vs. 'Vida social, diversões'; (c) Some entries are unique by design, that is the case with the themes of lectures and conferences; and (d) There are missing values in all features.

In this step, entries with small discrepancies were corrected and "no information" values were inserted in a new class "None" in the database and saved to be manually re-checked by the team. After the team revision, features were updated. Specifically for themes, a set of classes was defined taking the most common and representative from each theme text as a standard for the label. Also, a generic label for conferences and school classes was created.

4.6 Word Search

Storing text in a manner to facilitate word searches can be a challenge. Firstly, calling a script to search in real-time multiple text documents every time a user makes a word search demands server processing and increases the response time. The operation becomes slower and more demanding as the corpus size and the number of synchronically client requests grow.

Secondly, the high number of entries (words) and their relations with inquiries text (transcriptions) is an obstacle for RDB storage and queries. A preliminary survey towards MC characteristics (the smallest corpus in NURC-SP) showed more than 6000 unique words (words considered as sequences of letters split by spaces). Moreover, the amount of many-to-many relations would be considerably high, for instance, common words would be linked with most inquiries transcriptions. Consequently, for the Portal NURC-SP digital we chose not to have a table for inquiries' text, nor a vocabulary table in the RDB.

The solution adopted was to compute all vocabulary searches in advance and store them in a lookup table (O(1)). The response retrieves values from memory, instead of expensive repeated computation.

5 Final Remarks

We presented NURC-SP Digital and its web portal, an interactive repository to maintain, distribute, and organize the digital corpora from NURC-SP, the Sao Paulo division of the NURC Project. The corpora collection – MC, CATNA and AC – is the result of the TaRSila Project team to process, transcribe, and carefully revise NURC-SP original audios and transcriptions. A work that is mobilizing researches from multiple fields and involved the use of speech tools to make the processing faster and help human annotation. We designed and implemented the portal to attend the demands of multiple users, considering the need of programming and non-programming researchers, making use of knowledge from front-end and back-end programming, user interaction, interface design, data preprocessing, and database architecture.

That is the first release of the Portal NURC-SP Digital. It provides access to the NURC-SP Digital collection (audio, transcriptions and audio-text alignment files), a filtering mechanism for metadata of the inquiries (inquiry type, theme, year, speakers' range of age, gender, and recording quality and duration) and word search in transcriptions.

The Portal was publicly launched in December 15, 2023, and the status of the subcorpora processing[12] is the following. MC is fully annotated (automatically and manually revised) with its 21 inquiries inserted in the database of the Portal. CATNA has 12 inquiries with prosodic segmentation manually revised and the remaining 14 inquiries are in the revision process of the automatic prosodic segmentation. Regarding the AC, 328 inquiries have already been revised but five of them had their revision discarded because the audio files were too noisy. From this release, we will collect users' feedback to make oriented improvements.

Our next step is to learn from users' experience and interaction with the portal, including from the TaRSila Project team of linguists who intend to make use of the NURC-SP Digital Portal for Prosody's studies. Moreover, we will provide, in the NURC-SP Digital Portal, a release of the 323 inquiries of AC divided in train/dev/test partitions similar to the CORAA ASR dataset (`https://github.com/nilc-nlp/CORAA`) to evaluate speech recognition models in Brazilian Portuguese spontaneous speech. We believe that providing easy and organized access to the NURC-SP Digital collection will help future linguistic and computational research in the Portuguese spoken language domain.

[12] Data processing took place from December 2020 to December 2023.

Acknowledgements

First of all, we would like to thank the annotators of the TaRSila project who were tireless in reviewing the automatic transcriptions, training and testing the models for various speech processing systems. This work was carried out at the Artificial Intelligence Center (C4AI-USP), with support from the São Paulo Research Foundation (FAPESP grant nº 2019/07665-4) and IBM Corporation. We also thank the support of the Center of Excellence in Artificial Intelligence (CEIA) funded by the Goiás State Foundation (FAPEG grant no. 201910267000527), the São Paulo University Support Foundation (FUSP) and the National Council for Scientific and Technological Development (CNPq-PQ scholarship, process 304961/2021-3). This project was also supported by the Ministry of Science, Technology and Innovation, with resources from Law nº 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residência no TIC 13, DOU 01245.010222/2022-44.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, pages 4489–4493.

Eckhard Bick. 2000. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus, Århus.

Paul Boersma and David Weenink. 2023. Praat: doing phonetics by computer [Computer program]. Version 6.3.10.

António Branco, Amália Mendes, Paulo Quaresma, Luís Gomes, João Silva, and Andrea Teixeira. 2020. Infrastructure for the science and technology of language PORTULAN CLARIN. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 1–7, Marseille, France. European Language Resources Association.

Ken Chen and Mark Hasegawa-Johnson. 2004. How prosody improves word recognition. In *Proc. Speech Prosody 2004*, pages 583–586.

Giovana Meloni Craveiro, Vinícius Gonçalves Santos, Gabriel Jose Pellisser Dalalana, Flaviane R. Fernandes Svartman, and Sandra Maria Aluísio. 2024. Simple and fast automatic prosodic segmentation of brazilian portuguese spontaneous speech. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR*

2024), Santiago de Compostela, Galicia. Association for Computational Linguistics. To appear.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).

Cheng-Hsien Lin, Chung-Long You, Chen-Yu Chiang, Yih-Ru Wang, and Sin-Horng Chen. 2019. Hierarchical prosody modeling for Mandarin spontaneous speech. *The Journal of the Acoustical Society of America*, 145(4):2576–2596.

Shimeng Liu, Yoshitaka Nakajima, Lihan Chen, Sophia Arndt, Maki Kakizoe, Mark A. Elliott, and Gerard B. Remijn. 2022. How pause duration influences impressions of english speech: Comparison between native and non-native speakers. *Frontiers in Psychology*, 13.

Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

Miguel Oliveira Jr. 2016. NURC digital um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (NURC). *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 3(2):149–174.

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*, volume 202 of *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.

Tommaso Raso, Bárbara Teixeira, and Plínio Barbosa. 2020. Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. *Journal of Speech Sciences*, 9:105–128.

P.G. Rowe. 1991. *Design Thinking*. Mit Press. MIT Press.

Vinícius G. Santos, Caroline Adriane Alves, Bruno Baldissera Carlotto, Bruno Angelo Papa Dias, Lucas Rafael Stefanel Gris, Renan de Lima Izaias, Maria Luiza Azevedo de Morais, Paula Marin de Oliveira, Rafael Sicoli, Flaviane Romani Fernandes Svartman, Marli Quadros Leite, and Sandra Maria Aluísio. 2022. Coraa NURC-sp minimal corpus: a manually annotated corpus of brazilian portuguese spontaneous speech. In *Proc. IberSPEECH 2022*, pages 161–165.

Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

Luiz Antônio da Silva. 1996. Projeto NURC: Histórico. *Linha D'Água*, v1996(10):83–90.

Bárbara Teixeira, Plínio Barbosa, and Tommaso Raso. 2018. Automatic detection of prosodic boundaries in Brazilian Portuguese spontaneous speech. In *Computational Processing of the Portuguese Language*, pages 429–437, Cham. Springer International Publishing.

Bárbara Helohá Falcão Teixeira and Maryualê Malvessi Mittman. 2018. Acoustic models for the automatic identification of prosodic boundaries in spontaneous speech. *Revista de Estudos da Linguagem*, 26(4):1455–1488.