MuPe Life Stories Dataset: Spontaneous Speech in Brazilian Portuguese with a Case Study Evaluation on ASR Bias against Speakers Groups and Topic Modeling

Sidney Leal[1,6], Arnaldo Candido Jr.[2], Ricardo Marcacini[1], Edresson Casanova[3], Odilon Gonçalves[4], Anderson Soares[5], Rodrigo Lima[1], Lucas Gris[5], Sandra Aluísio[1]

[1]University of São Paulo, São Carlos, SP, Brazil
[2]Universidade Estadual Paulista, São José do Rio Preto, SP, Brazil
[3]NVIDIA Corporation, São Paulo, SP, Brazil
[4]Museu da Pessoa, São Paulo, SP, Brazil
[5]Centro de Excelência em Inteligência Artificial (CEIA-UFG), Goiânia, GO, Brazil
[6]Venturus - Centro de Inovação Tecnológica, Campinas, SP, Brazil
Correspondence: sandra@icmc.usp.br

Abstract

Recently, several public datasets for automatic speech recognition (ASR) in Brazilian Portuguese (BP) have been released, improving ASR systems performance. However, these datasets lack diversity in terms of age groups, regional accents, and education levels. In this paper, we present a new publicly available dataset consisting of 289 life story interviews (365 hours), featuring a broad range of speakers varying in age, education, and regional accents. First, we demonstrated the presence of bias in current BP ASR models concerning education levels and age groups. Second, we showed that our dataset helps mitigate these biases. Additionally, an ASR model trained on our dataset performed better during evaluation on a diverse test set. Finally, the ASR model trained with our dataset was extrinsically evaluated through a topic modeling task that utilized the automatically transcribed output.

1 Introduction

The research area on Automatic Speech Recognition (ASR) for Brazilian Portuguese (BP) started to become more active from mid-2020, with the availability of several public datasets for building such systems. The number of hours increased from 60 hours divided into four small datasets featuring read speech[1] to more than 1,500 hours in 2024 (see Table 1).

Brazil is a continental country, the largest and the most populous country located in South America, with 26 states and the Federal District, where the capital of Brazil (Brasília) is located, having many accents which are spoken across its five geographic regions (North, Northeast, Midwest, Southeast, and South). The Southeast is the region with the largest gross domestic product (GDP) in Brazil and that is why people from other regions migrate to this region. Migration also shows an aspect related to the education level in Brazil as one of the harshest aspects of social inequality in Brazil is the average level of education of the population, considered low in relation to other countries. The 2018 INAF (The National Indicator of Functional Literacy) report presented a worrying scenario: 8% of the individuals are illiterate and only 12% are literate at the proficient level[2]. Therefore, this scenario led us to study the influence of variables such as regional accents and education on the impact of ASR models for BP as the analysis of ASR bias was lacking for resources in BP presented in Table 1.

Although the NURC-SP Audio Corpus (Lima et al., 2025) is balanced in relation to the speakers' gender, there was no assessment of the impact of this variable on the trained ASR models. On the other hand, CORAA ASR (Candido_Junior et al., 2023) shows the performance of the model trained on the dataset in relation to regional accents and speaking styles (spontaneous vs read speech). The metrics Character Error Rate (CER) and Word Error Rate (WER) were used to assess model performance. The model trained on the entire dataset does not recognize regional accents equally well. For example, the performance of the ASR model on speech data from São Paulo state cities was the worst (34.06% of WER) and from São Paulo Capital the best (20% of WER), while the accents from Minas Gerais and Recife states had 28.88% and 22.03%, respectively of WER. With regard to speaking styles, the ASR model also performs differently, presenting 26.5% of WER for spontaneous speech and 18.7% for read speech. These results

[1] The Common Voice Corpus version 5.1 (commonvoice.mozilla.org/pt/datasets), Sid dataset, VoxForge, and LapsBM1.4 (igormq.github.io/datasets/)

[2] https://alfabetismofuncional.org.br/

Proceedings of the 31st International Conference on Computational Linguistics, pages 6076–6087
January 19–24, 2025. ©2025 Association for Computational Linguistics

are expected as speech recognizers perform worse with spontaneous speech audios[3]. However, there is a lack of research whether these differences are statistically significant when applied to read speech. This is true even in models trained with large volumes of data like Whisper (Radford et al., 2023). Regarding ASR models and overall WER results on trained BP datasets, Quintanilha et al. (2020) evaluates an end-to-end ASR system for Brazilian Portuguese with a DeepSpeech-2-based architecture (Amodei et al., 2016) and a transfer-learning approach from a backbone model trained on LibriSpeech dataset[4]. They used the BP speech dataset named BRSDv2 (see at Table 1) and built a new text corpus comprising 10.2 million sentences from the Portuguese Wikipedia. By combining a language model with an acoustic model, they achieved CER and WER values of 10.49% and 25.45%, respectively.

Gris et al. (2022) presents a fine-tuning of the Wav2vec 2.0 XLSR-53 model pre-trained in several languages (Conneau et al., 2021), over seven BP datasets totaling 470 hours of read and prepared speech (e.g. Common Voice 7.0 (Ardila et al., 2020), MultiLingual LibriSpeech (MLS) corpus (Pratap et al., 2020), MultiLingual TeDx Corpus (Salesky et al., 2021)). Their model achieved a WER value of 36.54% on average of all datasets (including spontaneous speech).

Candido_Junior et al. (2023) also present a public ASR model based on Wav2Vec 2.0 XLSR-53, fine-tuned over CORAA ASR corpus. CORAA ASR contains 290h of human validated pairs of audio-transcription from four BP spontaneous speech corpora and prepared speech from a set of TEDx Portuguese talks, a new corpus compiled specifically for CORAA ASR. Their model achieved a WER of 24.18% on CORAA ASR test set, and 29.44% on average of all datasets.

Differently from the previous work, Lima et al. (2025) report ASR results, using both the Wav2Vec2-XLSR-53 and Distil-Whisper models fine-tuned and trained on the NURC-SP Audio Corpus. This corpus is balanced regarding speaker's genders (204 females, 197 males) totaling 239.30 hours of transcribed audio recordings. Their best results were the Distil-Whisper fine-tuned over NURC-SP Audio Corpus with a WER of 24.22%, what motivated us to apply their distilled model for our study in this paper. Their model achieved 22.21% of WER on average in our evaluation.

In this paper, we present a freely available spontaneous speech corpus for the BP language, called **MuPe Life Stories**, composed of 289 life story interviews (135 females, 153 males), in 365 hours of automatically transcribed and manually revised audio recordings. Human transcription is time-consuming, especially for lengthy audios or videos, therefore an ASR dedicated to the domain of life stories methodology is very welcome. The dataset has a rich set of metadata information about interviewees. With this rich metadata we fine-tuned the public Distil-Whisper model presented at Lima et al. (2025) over MuPe Life Stories dataset and evaluate the ASR model regarding the biases that the ASR task has in relation to gender, age, regional linguistic varieties (accents from the 5 regions of Brazil) and education of the interviewees. Here, we follow Feng et al. (2024) which define bias as the difference in WER between the different speaker groups investigated; it is computed by subtracting the lowest WER from the WER of each speaker group analyzed. The main contributions made in this work are summarized as follows.

1. A large corpus of life story narratives for the task of ASR in BP, containing 365 hours of spontaneous speech. It is publicly available[5] under the CC BY-NC-ND 4.0 license.

2. To the best of our knowledge, we introduced the first corpus tackling speech recognition bias against gender, regional accents, age, and education level in BP.

3. We released publicly an ASR model[6] trained with our proposed dataset, which achieved SOTA results in PB, reducing biases and improving the ASR generalization.

4. We present a study on ASR-based video topic modeling [7], showing that topic models generated from ASR transcriptions using the Distil-Whisper model achieve quality comparable to manual transcriptions, despite transcription errors, supporting their practical use for

[3] Spontaneous speech has phenomena that make its recognition more complex than that of read or prepared speech, such as filled pauses and other disfluencies.

[4] `https://www.openslr.org/12`

[5] `https://huggingface.co/datasets/nilc-nlp/CORAA-MUPE-ASR`

[6] `https://huggingface.co/nilc-nlp/distil-whisper-coraa-mupe-asr`

[7] `https://github.com/nilc-nlp/coling-mupe-asr`

automatic organization of large-scale video datasets.

2 Related Work on BP Datasets for ASR Models Evaluation

Table 1 presents large corpora/datasets built specifically for ASR evaluation or assorted to be used in ASR evaluations. For multilingual resources, we show the statistics for the Portuguese language.

The MultiLingual LibriSpeech (MLS) corpus (Pratap et al., 2020) is a large multilingual corpus for speech research, derived from read audiobooks from LibriVox[8] and consists of 8 languages, including about 32 thousand hours of English and a total of 4.5 thousand hours of other languages. For Portuguese, there are 131 hours and 54 speakers. It is only suitable for speech recognition in low-noise scenarios.

BRSDv2 dataset (Quintanilha et al., 2020) is composed of CETUC dataset (Alencar and Alcaim, 2008) (designed by its authors for both synthesis and speech recognition applications) and four smaller datasets, three of them (Sid dataset, VoxForge, and LapsBM) freely available[9], totaling 158 hours of speech.

The MultiLingual TeDx Corpus (Salesky et al., 2021) was proposed to enable research in the areas of automatic speech recognition and speech-to-text translation. For Portuguese, there are 164 hours available in 93 thousand audios. The corpus is made up of talks on a wide range of subjects, being managed within the scope of the TEDx project, linked to the TED group.

CORAA ASR (Candido_Junior et al., 2023) is a corpus for ASR that contains specially spontaneous speech. CORAA ASR is the combination of five independent projects dealing with speech of the interior of São Paulo state (with 35.96 hours - ALIP Project (Gonçalves, 2019)), Minas Gerais state (with 9.64 hours - C-ORAL Brasil I Project (Raso and Mello, 2012)), Recife (capital of Pernambuco state) (141.31 hours - NURC-Recife Project (Oliveira Jr., 2016)) and São Paulo capital (31.14 hours - SP2010 Project (Mendes and Oushiro, 2012)), in addition to the prepared speech of TeDx Talks in Brazilian and European Portuguese (72.74 hours), totaling 290 hours.

Common Voice corpus (Ardila et al., 2020) is an open-use project created by the Mozilla Foundation were users can simultaneously contribute to the growth of the base and access other people's audio. To collaborate with the project, users can donate audio in their own voices and review donations from other users. The permissive-use license of this project allows the exploration of the corpus including for commercial purposes. In version 17, launched in 2024, the subcorpus for the Portuguese language has 211 hours of audio and transcriptions, of which 175 were validated.

NURC-SP Audio Corpus (Lima et al., 2025) is a freely available spontaneous speech corpus for the BP language and comprises 401 different speakers (204 females, 197 males) with a total of 239.30 hours of transcribed audio recordings. It is a large Paulistano accented spontaneous speech corpus dedicated to the ASR task in Portuguese. Despite the small numerical difference, the NURC-SP Audio Corpus is actually less diverse than MuPe Life Stories, because NURC-SP Audio Corpus only has speakers with higher education and a São Paulo city accent.

The MuPe Life Stories Dataset, the focus of this paper, is described in detail in Section 3.

3 The MuPe Life Stories Dataset

The MuPe Life Stories Dataset is composed of **256** life story narratives taken from the "Projeto 25 anos de Museu da Pessoa no Brasil" (25 years of the Museum of the Person in Brazil project) (Section 3.1) and **33** life story narratives from the "Projeto Memórias dos Brasileiros" (Brazilian Memory Project) (Section 3.2), totaling 365.15 valid hours.

The dataset has a rich set of metadata information, including interviewer's names, title of the interview, interviewee name and gender, birthday, religion, education, birth city, state and county, occupation, profession, besides a short biography. The interviewees were all born among 1905 and 1991 and for the majority of them, there is information about the state and country where they were born. Figure 3 (Appendix B) shows the distribution of interviewees by state of birth; there is great discrepancy regarding state of origin, where São Paulo state stands out with 167 life stories. There are 17 speakers who were born in different countries, such as Portugal, Chile, Japan, Germany, Italy, but live in Brazil and speak Portuguese.

These 289 life stories in total were collected by the virtual and collaborative museum named *Museu*

[8] https://librivox.org/

[9] https://igormq.github.io/datasets/

Corpora	Hours	Speaking Style	Number of Speakers	License
MultiLingual LibriSpeech (MLS) (Pratap et al., 2020)	130.1	read	54	CC BY
BRSD v2 (Macedo Quintanilha et al., 2020)	158	read	775	-
Multilingual TEDx (Salesky et al., 2021)	164	prepared	-	CC BY-NC-ND 4.0
CORAA ASR 1.1 (Candido Junior et al., 2023)	290	spontaneous	1,689	CC BY-NC-ND 4.0
Common Voice 17.0 (Ardila et al., 2020)	175	read	3,453	CC-0
NURC-SP Audio Corpus (Lima et al., 2024)	239.30	spontaneous	401	CC BY-NC-ND 4.0
MuPe Life Stories Dataset (ours)	365	spontaneous	289	CC BY-NC-ND 4.0
Total	1,521.4			

Table 1: Statistics of the public datasets available for ASR in Portuguese.

da Pessoa[10] (Museum of the Person). A cooperation agreement between Museum of the Person and two Brazilian universities was formalized to allow the building of MuPe Life Stories Dataset. The life stories were processed and anonymized to be part of our dataset (Section 3.3).

3.1 25 Years of the Museum of the Person in Brazil Project

There are 280 life stories in this project, most of them with content available (several including transcription and full video available) and revised on the web platform of Museum of the Person (MuPe), which were digitized and edited as part of the "25 years of the MuPe in Brazil: Strengthening and Consolidation of Assets"[11] project.

The collection of the MuPe is made up of life stories, told by the people themselves or by third parties. The narratives are recorded in three ways: (i) at the Museum's headquarters, in a studio — recorded on video and collected by interviewers trained on life history methodology, (ii) sent via the internet by *Programa Conte sua História (Tell Your Story Program)* or (iii) via *Museu que Anda (Walking Museum)*, a program in which the narratives of people outside the headquarters are recorded through itinerant booths. Each interview constitutes a unit of the collection that is formed by the audio or video recording of the interview, the transcription and edition of each narrative, accompanied by photos and documents sent by the people who tell their life stories.

Once recorded, life stories collected by MuPe are transcribed and sometimes revised. The transcripts have annotations of laughter, clapping hands, whistles, emotional speech, pauses, among others, using parentheses. Also, expansions of acronyms are annotated using square brackets. Moreover, the transcription is segmented and a proposal of punctuation is provided using seven punctuation marks (Exclamation, FullStop and Question Marks, Ellipsis, Comma, Semicolon and Colon). The turns are indicated by P/1 (and P/2 and P/3) and R labels followed by the transcription of the turn, where P/i (i = 1, 2 or 3) indicates the interviewer and R the interviewee. However, since disfluencies (corrections and repetitions) that are common in spontaneous speech are not annotated, the MuPe transcriptions can be called an adapted verbatim transcription. Therefore, in order to create the MuPe Life Stories Dataset we decided to generate a verbatim transcription using WhisperX (Bain et al., 2023) followed by manual revision of the automated transcription (see Section 3.3 for details).

3.2 Brazilian Memory Project

The Brazilian Memory Project project[12] began to be compiled between the years 2003/2004. At that time, close to completing 10 years of operation, the Museum of the Person began a process of redirecting its activities, starting to work more with communities, organizations and social groups. The project was born from the idea that it was necessary to record stories of invisible people, living in a country that was undergoing complete transformation — the Brazilian Memory Project was then created and the life stories were collected through a series of expeditions carried out since then and today the collection already exceeds 300 life stories. Throughout the development of the project, organizations were interested in joining efforts and resources to enable a broad recording of life stories across different thematic lines. These collaborative partnerships allowed for the richness of the narratives and the diversity of the stories recorded. By giving voice, eyes and ears to the life story narratives, images, works and manifestations of people from different origins, trajectories and realities, the

[10] https://museudapessoa.org/historias/

[11] https://acervo.museudapessoa.org/pt/apoie/quem-apoia/apoio-bndes

[12] https://museudapessoa.org/acoes/memoria-dos-brasileiros-guia-do-acervo/

project intended to collaborate with overcoming prejudices and expanding and deepening understanding of current conflicts and the paths of the future. The project helped bring the Internet's general public closer to various social actors spread across the various "Brazils" that coexist within Brazil, by providing a new source of knowledge for Education, Media, Science, Art, Public Policies, among others.

3.3 Pre-processing, Anonymization and Normalization for Evaluating ASR Models

The life stories were automatically transcribed by WhisperX (Radford et al., 2023) which provides fast transcription using the large-v2 model of Whisper and diarization via pyannote-audio[13].

Ten trained students revised the automatic transcriptions from June 2023 to April 2024. The revision process was based on an annotation guideline designed to: (i) help making the revision uniform and (ii) remove segments with loud noise, overlapping voices and, mainly, people's names (interviewee's name, father's and mother's names, grandfather's and grandmother's names) in order to comply with the data session rules of the agreement with the Museum of the Person. The guideline contains 11 rules, dealing with (i) orality marks, (ii) how to transcribe filled pauses, (iii) repetitive hesitations, (iv) numbers, (v) individual letters, (vi) acronyms, (vii) foreign terms, (viii) punctuation and capitalization, (iv) emotion sounds (ex. laughter) which were annotated in parentheses, (x) misunderstanding of words or passages and (xi) how to deal with automatic segmentation failures. The evaluation of pyannote-audio diarization labels was carried out in May 2024 with two trained students, throughout a long life history which was divided into 1,146 segments by WhisperX and Cohen's Kappa for the 2 raters was 0.947, considered almost perfect (Landis and Koch, 1977). The revised transcriptions of MuPe Life Stories Dataset use upper and lower case letters and punctuation, as well as, filled pause markers such as *eh*, *hum*, *ãh*, etc.

3.4 Statistics of MuPe Life Stories Dataset

This new dataset has more than 365 hours divided into training, validation, and test subsets, with a total of 317,746 audio segments (average of 4.14 seconds). Table 2 presents detailed information about the MuPe Life Stories training, validation, and test sets after the pre-processing steps. The training set is the biggest set, containing approximately 87% of the total hours (∼68% of life stories), while the validation set has ∼3% (both in hours and life stories) and test set contains 9% and 10% of audios and life stories, respectively. The three subsets are well balanced in terms of gender, although the division is not perfect, as we tried to balance regions equally as well. Therefore, there are slightly more males in the training set and females in the validation and test sets. The dataset contains more than 3.6M tokens, with broad vocabulary coverage, given the type/token rate of ∼87%.

Table 9 (Appendix B) shows seven metadata information of the MuPe Life Stories testset used in the Bias Study. The selection of life stories to compose the testset was carried out in the work of Craveiro and Galdino (2025) which highlighted the importance of considering different speaker profiles to meet the ethical principle of AI diversity and, thus, promote technology that is equally efficient for all types of speakers. The column education had some information completed after we received the data from Museum of the Person. Although information about gender, state of origin and year of birth are almost completed in dataset metadata (99%), only 21% of education level are indicated for life stories of training and validation subsets.

4 Baseline Model Development

We performed an ASR experiment over MuPe Life Stories Dataset in order to measure the dataset quality, potentials, and limitations.

Before the execution of this experiment, we normalized train, validation, and test subsets in order to simplify the training and calculation of the CER and WER metrics. The following normalization was performed: (i) the texts were transformed into lowercase; (ii) all punctuation marks generated by Whisper were removed (ellipsis, exclamation mark, fullstop, question mark, and comma); (iii) the filled pauses were standardized to: *eh, uh, ah*, as follows: *eh* = *eh, éh, ehm, ehn*; *uh* = *uh, hm, uhm, hmm, mm, mhm*; *ah* = *ah, huh, ãh, ã*; and (iv) successive blank spaces have been replaced with one space. It is important to note that this process occurred as a post-processing step, only for the evaluation of CER and WER metrics.

For our experiments, we followed the work by Lima et al. (2025) which used Distil-Whisper

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

	Training	Validation	Test	Total
Number of Life Stories	250	9	30	289
Female voices (number)	114	5	16	135
Male voices (number)	135	4	14	153
Male/Female Ratio	1.18	0.80	0.87	1.13
Duration (hours)	319.99	12.25	32.91	365.15
Qty Audios (segmented)	276,881	9,894	30,971	317,746
Segment Duration (avg seconds)	4.16	4.46	3.83	4.14
Segment Duration (max seconds)	29.81	29.41	29.65	29.81
Avg Tokens	11.53	12.18	11.05	11.50
Avg Types	10.10	10.57	9.69	10.07
Total Tokens	3,192,306	120,535	342,348	3,655,189
Total Types	2,796,532	104,567	299,959	3,201,058
Type/Token Ratio	0.876	0.868	0.876	0.876

Table 2: Statistics of MuPe Life Stories Dataset.

Datasets	Common Voice		CORAA ASR		NURC-SP Audio Corpus		MuPe Life Stories		Mean	
Metrics	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
(Gris et al., 2022)	**4.50**	**16.32**	22.32	43.70	26.52	47.74	18.57	38.38	17.98	36.54
(Candido_Junior et al., 2023)	6.99	24.44	**11.02**	**24.18**	22.87	40.29	14.04	28.84	13.73	29.44
(Lima et al., 2025)	5.70	17.76	14.89	26.91	**15.77**	**24.22**	11.54	19.94	11.98	22.21
Ours	5.17	16.38	14.19	26.14	17.32	26.58	**9.56**	**15.99**	**11.56**	**21.27**

Table 3: MuPe Life Stories ASR model results compared to previous works.

(Gandhi et al., 2023) as ASR model. We used the pre-trained model released by Lima et al. (2025) and fine-tuned it using our dataset. This pre-trained checkpoint[14] was selected because it was already optimized for BP on a dataset featuring spontaneous speech with the accent spoken in the capital of São Paulo state, which is also the prevalent accent in the MuPe Life Stories Dataset (see Figure 3 in Appendix B). The model was then trained for more 18k steps and batch size of 16, for approximately 50 hours in an Nvidia L4 Tensor Core GPU inside Google Cloud infrastructure.

4.1 Results and Discussions

To evaluate the model trained on our dataset, we compared it with previous SOTA BP models (Gris et al., 2022; Candido_Junior et al., 2023; Lima et al., 2025) across multiple datasets, as shown in Table 3.

In general, all models tend to perform better on the datasets they were trained on. For instance, Gris et al. (2022) achieved superior results on the Common Voice dataset, which was part of their training data. Similarly, the model by Candido_Junior et al. (2023) performed best on their CORAA dataset, and Lima et al. (2025) showed optimal performance on their proposed dataset. Our model also achieved the best results on our proposed dataset (MuPe Life Stories). Notably, our model ranked second best across all other datasets and showed superior performance when averaged across all datasets. These results suggest that our dataset enhances the generalization of the ASR model across all evaluation sets, highlighting the quality of our proposed dataset.

5 Bias in ASR for BP Spontaneous Speech

In Section 4, we demonstrated that the model trained with our proposed dataset achieved better generalization compared to previous related works. We attribute this improvement to the increased diversity of our dataset. To validate this hypothesis, we conducted a detailed bias analysis.

For the bias analysis, we have used our test set and we grouped the speakers by (i) Age, with intervals from 0 to 40, 40 to 60, and above 60; (ii) by Education Level, in three groups — elementary school/no education, high school/technical level and bachelor's degree/masters' degree; (iii) by Gender (male/female); and (iv) by Regional Accents grouped by geographic regions and clus-

[14] github.com/nilc-nlp/nurc-sp-audio-corpus

tered by proximity: North/Midwest, Northeast and South/Southeast. To determine whether significant differences in bias exist among the proposed groups, we applied a One-Way ANOVA, based on the results of the previous model trained by Lima et al. (2025) and our proposed model. For more details, please check the Appendix A. The results of bias analysis are discussed in Section 5.1.

5.1 Bias Evaluation

Table 4 presents a comparison between our model and the model proposed by Lima et al. (2025) across four target bias categories: Education Level, Age, Regional Accents, and Gender. Table 4 summarizes the data presented in Tables 7 and 8 (Appendix A).

The average WER was calculated by averaging the WER across all subcategories within each bias class as shown in Figure 2 (Appendix A). For instance, in the Education level category, the WER was averaged across all education levels in the subgroups Bachelor's Degree, Elementary School, High School. Additionally, the inter-class WER standard deviation (SD) was computed for each bias class to assess variability. To further evaluate the extent of this variability in relation to WER, the relative standard deviation was also calculated, representing the SD as a percentage of the WER. Finally, p-values for the One Way ANOVA were also reported. Considering the p-values, it can be observed that there is a significant difference between Education levels and Age groups for both our model and the model trained by Lima et al. (2025). However, despite the large WER gap between the North/Midwest and South/Southeast regions, as shown in Tables 7 and 8, no significant difference was found for Regional Accents. Similarly, no significant differences were observed between different Gender groups.

An analysis of relative standard deviations indicates that the model trained with our proposed dataset successfully reduced the WER gap between Education levels and Age groups, demonstrating that our dataset contributed to bias mitigation. Although the One-Way ANOVA indicated that Regional Accent bias is not statistically significant, we observed an improvement in the relative standard deviation from 15.73% to 11.32%. A small improvement was observed for the Gender group.

6 Experiments on Topic Modeling

In this work, we also explore topic modeling for videos using the ASR output of the life story interviews. This task supports the thematic organization of videos and video segments based on their textual transcriptions. We evaluated the topic modeling task on the MuPe Life Stories dataset by comparing topics generated from manually created transcriptions with those derived from ASR outputs. Our analysis aimed to assess whether transcription errors (e.g. word error rate) significantly affect the formation of topics. This evaluation also considered Age and Gender of the interviewees to determine if these factors influence the quality and accuracy of the topic modeling results.

Table 5 illustrates an example of how topics are associated with segments of video transcriptions. A set of top-k topics is initially obtained for the entire video, considering the global context, and then each segment is assigned a specific topic. This allows filtering of segments by predefined topics and supports tasks such as interview summarization based on topics of interest. The goal is for the automatically obtained topic organization to align with a reference organization (labeled topics).

For the experiments with video ASR-based topic modeling, we used the same test set presented in Tables 2 and 9. The reference topic organization was constructed based on manual transcriptions. Topics were automatically obtained both from the manual transcriptions and from the automatic transcriptions generated by Distil-Whisper trained in our dataset. We used the Normalized Mutual Information (NMI) measure to compare the obtained topics (Kvålseth, 2017). In this scenario, NMI allows for the comparison of automatically obtained topics with a reference topic organization.

We employed an ASR-based topic modeling approach leveraging Large Language Models (LLMs), which has shown promising results for topic modeling (Wang et al., 2023, 2024). In this study, we used Llama 3.1 8B Instruct model, with 8 billion parameters, publicly available[15](Dubey et al., 2024), in a zero-shot learning setting and in a two-step process. In the first step, the complete transcriptions of a video are presented to the LLM with a prompt to generate and enumerate the main topics identified in the interview, thereby obtaining a set of top-k topics, where k is determined automatically. In the second step, transcription

[15] https://ollama.com/library/llama3.1

	model	**Avg WER**	**SD**	**Relative SD**	**p-value**
Education Level	Ours	15.56	3.70	**23.74**	<0.01
	Lima et al	19.37	4.98	25.70	<0.01
Age	Ours	15.53	3.28	**21.10**	<0.01
	Lima et al	19.52	4.41	22.58	<0.01
Regional Accents	Ours	16.30	1.84	**11.32**	0.26
	Lima et al	20.87	3.28	15.73	0.11
Gender	Ours	16.15	0.92	5.70	0.41
	Lima et al	20.41	1.08	**5.27**	0.47

Table 4: Bias comparison between Lima et al. (2025) model and ours.

Start Time	**End Time**	**Transcription Text**	**Labeled Topic**	**Extracted Topic**
00:00:01	00:00:15	"My name is João, I was born in São Paulo and always worked in the countryside."	Life Story	Personal Biography
00:00:16	00:01:00	"During the 1980s, the job market was very difficult for farmers."	Rural Economy	Economic Challenges
00:01:01	00:01:45	"Education in rural areas was always very poor, and many children had no access."	Education in Rural Areas	Education and Inequality
00:01:46	00:02:30	"In 1990, I started noticing major changes in agriculture with new technologies."	Technological Advances in Agriculture	Technological Innovations
00:02:31	00:03:15	"The government implemented several public policies to help small farmers."	Public Policies and Agriculture	Government Policies
00:03:16	00:03:50	"Many farmers still struggle with adapting to new technologies, even today."	Technological Advances in Agriculture	Technological Innovations
00:03:51	00:04:30	"Today, things are a little better, but the struggle continues for many small farmers."	Current Challenges for Small Farmers	Agriculture and Current Challenges

Table 5: Example of topic modeling for video transcriptions with repeated topics.

segments are presented again to the LLM with instructions to allocate each segment to one of the topics obtained in the previous step. In this second stage, we only used segments containing more than 20 tokens to eliminate very short segments that included only introductions, interjections, or segments with no significant semantic content.

Figure 1 presents a comparison chart of the NMI measure (y-axis) for ASR-based topic modeling using both manual and automatic transcriptions across 30 videos from the test set (x-axis). We observed that the NMI remained similar in both scenarios. Practically, this suggests that our ASR model can effectively be used for video topic modeling based on automatic transcription, without significantly reducing the performance of topic modeling compared to what would be achieved with manual transcriptions.

In Table 6, we show the impact of Age and Gender of the interviewees on topic formation. Across all these factors, the topics remained consistent when comparing topics obtained from manual transcriptions and automatic transcriptions. It is important to note that the LLM has built-in mechanisms to correct potential transcription errors based on context. We found that most errors occurred in terms that did not significantly alter the overall meaning of the segments (mainly in longer segments with more than 20 tokens).

	Manual Transcription	**Automatic Transcription**
Gender		
Male	0.226 ± 0.126	0.218 ± 0.121
Female	0.218 ± 0.108	0.199 ± 0.102
Year of birth (Age)		
$[1900, 1945)$	0.286 ± 0.133	0.265 ± 0.129
$[1945, 1967)$	0.164 ± 0.075	0.156 ± 0.072
$[1967, 1987)$	0.194 ± 0.083	0.184 ± 0.081

Table 6: Impact of age and gender of the interviewees on topic formation.

7 Conclusions

In this paper we presented a large BP corpus of revised audio-transcription pairs, publicly available, containing 365 hours of spontaneous speech. Moreover, we presented a Distil-Whisper model based

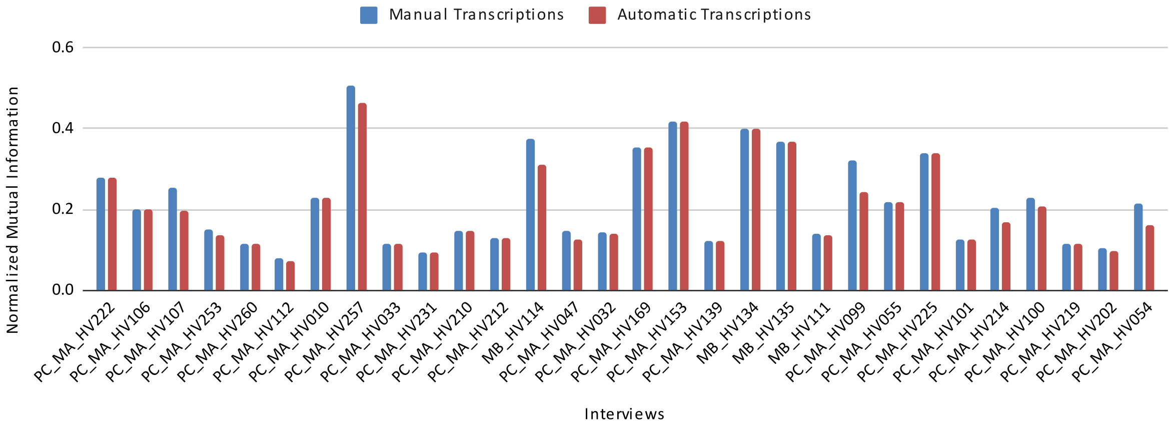

Figure 1: Comparison of ASR-based topic modeling using both manual and automatic transcriptions.

on the presented dataset that consistently perform as the best or second best model in BP evaluation datasets used in this work. In addition, we observed ASR bias regarding age and education levels, while no significant differences were found for gender and regional accents of Brazil.

The results indicate that the speech of people in the age group (40-60) was better recognized than the speech in the age group (0-40). The speech of senior speakers (60+) resulted in the worst performance of the ASR model. The speech of the bachelor's degree and superior degrees group was recognized better than those from high school and technical level group. The speech of the group with no education and elementary school resulted in the worst performance of the ASR. No bias was observed for females compared to the speech of males, although the speech of females showed better WER results. The regional accents of the South/Southeast group showed better WER results, although no significant differences were found in relation to the Northeast group or the North/Midwest group.

8 Limitations and Future Work

Several metadata information is missing in our dataset[16]. In future work, we intend to use the application of topic modeling to easily complete these missing information. Human transcription is time-consuming, and even revising automatically transcribed audio is an expensive task. The revision work done in MuPe Life Stories took 10 months, and was carried out by 10 trained annotators. In order to enlarge the dataset to balance the spoken accents (see Figure 3 in Appendix B) we intend to apply named entity recognition (NER) in BP as the largest number of Whisper errors are related to names of people, places and organizations in Portuguese. This will allow the training of BP models with accent-based batch balancing (Darshana et al., 2022; Maison and Esteve, 2023). The 17 life stories of people born in different countries are in the training dataset, but they would be better allocated in the test dataset to allow the bias evaluation of non-native Portuguese speakers in the dataset. This evaluation was carried out in the work by Feng et al. (2024) which finds speech recognition bias related to gender, age, regional accents and non-native accents for Dutch and Mandarin. There was no analysis of transcription errors for the model developed in this work, to verify whether errors related to NER, present in the Whisper transcriptions, reduced in the new model; this will be carried out in future work.

9 Acknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. This project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law No. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44. This work has been partially supported by Advanced Knowledge Center in Immersive Technologies (AKCIT/CEIA), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI, signed with EMBRAPII.

[16] https://github.com/nilc-nlp/coling-mupe-asr

References

V. F. S. Alencar and A. Alcaim. 2008. LSF and LPC - derived features for large vocabulary distributed continuous speech recognition in Brazilian Portuguese. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1237–1241.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

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.

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

Arnaldo Candido_Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio. 2023. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. *Lang Resources & Evaluation*, 57:1139–1171.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Giovana Meloni Craveiro and Julio Cesar Galdino. 2025. Diversity in data for speech processing in Brazilian portuguese. In *Intelligent Systems: 34th Brazilian Conference, Bracis 2024, Belém Do Pará, Brazil, November 17-21, 2024, Proceedings, Part I*. Springer Nature Switzerland.

S Darshana, H Theivaprakasham, G Jyothish Lal, B Premjith, V Sowmya, and Kp Soman. 2022. Mars: A hybrid deep cnn-based multi-accent recognition system for english language. In *2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*, pages 1–6.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Sebastião Carlos Leite Gonçalves. 2019. Projeto ALIP (amostra linguística do interior paulista) e banco de dados iboruna: 10 anos de contribuição com a descrição do português brasileiro. *Revista Estudos Linguísticos*, 48(1):276–297.

Lucas Rafael Stefanel Gris, Edresson Casanova, Frederico Santos de Oliveira, Anderson da Silva Soares, and Arnaldo Candido Junior. 2022. Brazilian Portuguese speech recognition using wav2vec 2.0. In *Computational Processing of the Portuguese Language*, pages 333–343, Cham. Springer International Publishing.

Tarald O Kvålseth. 2017. On normalized mutual information: measure derivations and properties. *Entropy*, 19(11):631.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Rodrigo Lima, Sidney Leal, Arnaldo Candido Jr., and Sandra Maria Aluísio. 2025. A large dataset of spontaneous speech with the accent spoken in são paulo for automatic speech recognition evaluation. In *Intelligent Systems: 34th Brazilian Conference, Bracis 2024, Belém Do Pará, Brazil, November 17-21, 2024, Proceedings, Part I*. Springer Nature Switzerland.

Lucas Maison and Yannick Esteve. 2023. Improving accented speech recognition with multi-domain training. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ronald Beline Mendes and Livia Oushiro. 2012. Mapping paulistano portuguese: the sp2010 project. In *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*, pages 459–463, Firenze, Italy. Fizenze University Press.

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.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Igor Macedo Quintanilha, Sergio Lima Netto, and Luiz Wagner Pereira Biscainho. 2020. An open-source end-to-end asr system for Brazilian Portuguese using DNNs built from newly assembled corpora. *Journal of Communication and Information Systems*, 35(1):230–242.

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 Machine Learning Research*, pages 28492–28518. PMLR.

Tommaso Raso and Heliana Mello. 2012. *C-oral - Brasil I: Corpus de Referência do Português Brasileiro Falado Informal*. Editora UFMG, Belo Horizonte, MG.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proc. Interspeech 2021*, pages 3655–3659.

Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. 2023. Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241. IEEE Computer Society.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2024. Large language models are latent variable models: explaining and finding good demonstrations for in-context learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Bias Results

The results show that WER becomes progressive smaller the higher the education level is (19.58%, 16.45%, and 10.66% for elementary, high school and higher education, respectively). This behavior is possibly explained due to the use of pre-training in our model. Our model is based on Lima et al. (2025), which, in turn, was more exposed to audios from speakers with higher education. Another possible reason of this behavior is that bachelors may use a more diverse vocabulary, adhere more frequently to the cultured norm and can expresses thoughts and reasoning in a more fluent manner.

Regarding age groups, the relation in which age affects ASR performance is less direct. WER is lower for the age groups from the interval 40-60. Regarding older people, the performance is probably explained by education level. From the test dataset presented in Table 9, no one above 60 year have completed higher education. The same occurs in the training set, as literacy improved during the 20th century as a result of public policies and the rural exodus in Brazil. Further investigation is required to explain the relatively low WER for the youngest group, as there is no visible education gap between this group and the interval 40-60.

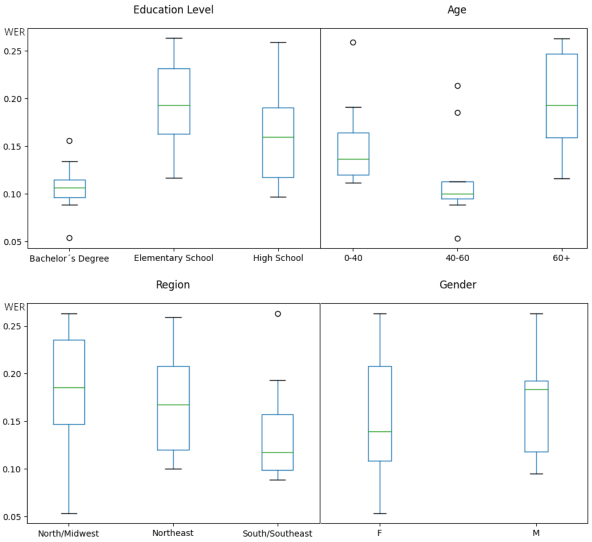

Figure 2: WER's comparison in all groups.

There was a non-significant difference among regional accents. However, it should be noted that the WER values for South and Southeast are lower than the other regions. This is possibly related to either accent differences, some presenting themselves easier and others harder to the ASR model, or educational levels differences, as South and Southeast are the most economically developed regions in Brazil, facilitating education access.

Likewise, there was a non-significant difference between gender, although WER values for female voices were slight smaller (∼2%). These results suggested that the training and pre-training data is roughly evenly distributed over genders.

Group	N	Mean	SD	SE	F-value	P-value
Bachelor's Degree	10	0.1066	0.0272	0.0086		
Elementary School	15	0.1958	0.0494	0.0128	11.2789	0.00027
High School	5	0.1645	0.0641	0.0287		
0-40	8	0.1534	0.0499	0.0176		
40-60	9	0.1162	0.0504	0.0168	6.6921	0.00436
60+	13	0.1964	0.0521	0.0144		
North/Midwest	7	0.1811	0.0758	0.0287		
Northeast	12	0.1703	0.0545	0.0157	1.3977	0.26447
South/Southeast	11	0.1377	0.0530	0.0160		
F	16	0.1523	0.0625	0.0156	0.6916	0.41265
M	14	0.1707	0.0582	0.0156		

Table 7: One Way ANOVA applied to all groups. (N = Number of different speakers, SD = Standard Deviation, SE = Standard Error).

Group	N	Mean	SD	SE	F-value	P-value
Bachelor's Degree	10	0.1318	0.0323	0.0102		
Elementary School	15	0.2537	0.0725	0.0187	11.9934	0.00018
High School	5	0.1956	0.0668	0.0299		
0-40	8	0.1850	0.0497	0.0176		
40-60	9	0.1471	0.0654	0.0218	6.8442	0.00394
60+	13	0.2536	0.0789	0.0219		
North/Midwest	7	0.2488	0.1191	0.0450		
Northeast	12	0.2089	0.0573	0.0165	2.3508	0.11450
South/Southeast	11	0.1684	0.0640	0.0193		
F	16	0.1933	0.0896	0.0224	0.5160	0.4784
M	14	0.2148	0.0715	0.0191		

Table 8: One Way ANOVA applied to all groups with Results from the previous best model of the literature (Lima et al., 2025). (N = Number of different speakers, SD = Standard Deviation, SE = Standard Error).

B MuPe Life Stories Test Subset: Metadata and Accents

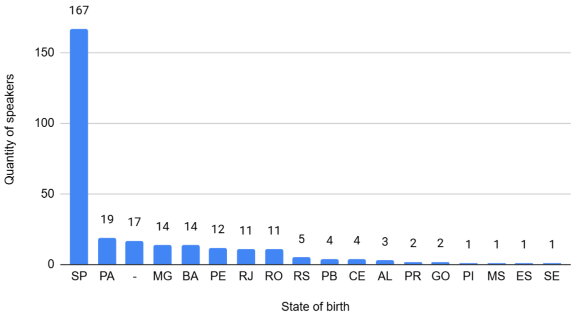

Figure 3: MuPe Life Stories Dataset distribution of speakers by state of birth.

In Figure 3, SP stands for São Paulo state; PA for Pará, MG for Minas Gerais, PE for Pernambuco, RJ for Rio de Janeiro, RO for Rondônia, RS for Rio Grande do Sul, PB for Paraiba, AL for Alagoas, GO for Goias, PI for Piaui, MS for Mato Grosso do Sul, ES for Espírito Santo, and SE for Sergipe. The speakers labeled with "-" were born in different countries, such as Portugal, Chile, Japan, Germany, Italy, but live in Brazil and speak Portuguese.

MuPe Code	G	Region/State	City	YOB/Age	Education
PC_MA_HV032	M	SUL/RS	Rio Grande	1952 (55 yrs)	Incomplete Bachelor's degree
PC_MA_HV212	F	SUL/RS	Bagé	1987 (22 yrs)	Incomplete Bachelor's degree
PC_MA_HV210	M	SUL/PR	Rolândia	1946 (63 yrs)	Technical Level
PC_MA_HV260	F	SUL/PR	Curitiba	1969 (41 yrs)	Complete Bachelor's degree
PC_MA_HV257	M	CENTRO-OESTE/GO	Catalão	1937 (73 yrs)	Complete Elementary School
PC_MA_HV033	M	CENTRO-OESTE/GO	Goiânia	1950 (57 yrs)	Complete Bachelor's degree
PC_MA_HV054	F	CENTRO-OESTE/MS	Cassilândia	1965 (42 yrs)	Complete Bachelor's degree
MB_HV134	F	NORTE/RO	Mutum-Paraná	unk (2010)	No education level
MB_HV135	M	NORTE/RO	Jaci-Paraná	1943 (67 yrs)	No education level
MB_HV111	F	NORTE/PA	Juruti	1927 (83 yrs)	No education level
MB_HV114	M	NORTE/PA	Castanhal	1962 (48 yrs)	Incomplete Elementary School
PC_MA_HV225	F	NORDESTE/PI	Teresina	1965 (45 yrs)	Complete Bachelor's degree
PC_MA_HV169	F	NORDESTE/SE	Simão Dias	1981(27 yrs)	Technical Level
PC_MA_HV099	F	NORDESTE/AL	Cacimbinhas	1940 (68 yrs)	No education level
PC_MA_HV253	M	NORDESTE/AL	Santana de Ipanema	1942 (68 yrs)	Incomplete Elementary School
PC_MA_HV139	F	NORDESTE/PB	Serra Branca	1945 (63 yrs)	No education level
PC_MA_HV010	M	NORDESTE/PB	Imaculada	1967 (40 yrs)	No education level
PC_MA_HV106	M	NORDESTE/CE	Tauá	1971 (37 yrs)	Incomplete Elementary School
PC_MA_HV100	F	NORDESTE/CE	Várzea Alegre	1952 (56 yrs)	Incomplete Elementary School
PC_MA_HV214	F	NORDESTE/BA	Abaíra	1925 (84 yrs)	No education level
PC_MA_HV107	M	NORDESTE/BA	Lençóis	1978 (30 yrs)	Incomplete Elementary School
PC_MA_HV202	F	SUDESTE/ES	Vitória	1965 (44 yrs)	Incomplete Bachelor's degree
PC_MA_HV055	F	SUDESTE/RJ	Rio de Janeiro	1936 (71 yrs)	Incomplete high school
PC_MA_HV231	M	SUDESTE/RJ	Duque de Caxias	1965 (44 yrs)	Technical Level
PC_MA_HV219	M	SUDESTE/MG	Minas Novas	1943 (66 yrs)	No education level
PC_MA_HV112	F	SUDESTE/MG	Ibiá	1968 (40 yrs)	Complete Bachelor's degree
PC_MA_HV153	M	NORDESTE/PE	Caruaru	1917 (91 yrs)	Technical Level
PC_MA_HV047	F	NORDESTE/PE	Manari	1987 (20 yrs)	Incomplete Bachelor's degree
PC_MA_HV222	M	SUDESTE/SP	Catanduva	1922 (88 yrs)	No education level
PC_MA_HV101	F	SUDESTE/SP	Mogi das Cruzes	1981 (27 yrs)	Master's degree

Table 9: Test Dataset with information about each life story (first column presents the Story ID in our dataset) used in the Bias Study. The table contains information about sex (S) (M:male, F:female), region/state (SP stands for São Paulo state; PA for Pará, MG for Minas Gerais, PE for Pernambuco, RJ for Rio de Janeiro, RO for Rondônia, RS for Rio Grande do Sul, PB for Paraiba, AL for Alagoas, GO for Goias, PI for Piaui, MS for Mato Grosso do Sul, ES for Espírito Santo, and SE for Sergipe) and city of origin, year of birth (YOB), and the education level of the interviewee. Given the nature of this data, which was collected through the interview, one information about YOB is lacking, which is marked with "unk" (date of the interview) for "unknown".