Roda Viva boundaries: an overview of an audio-transcription corpus

Isaac Souza de Miranda Jr.
Federal University of
São Carlos / Brazil
isc_jr@live.com

Gabriela Wick-Pedro
Center for Artificial Intelligence / Brazil
gabiwick@gmail.com

Cláudia Dias de Barros
Federal Institute of Education,
Science and Technology of São Paulo / Brazil
claudias84@gmail.com

Oto Vale
Federal University of
São Carlos / Brazil
otovale@ufscar.br

Abstract

This paper highlights the initial steps and challenges in creating an audio transcription corpus focused on interviews, intended for linguists, computer scientists, historians, sociologists, political scientists, communication professionals, and digital humanities researchers. The Roda Viva corpus, derived from the renowned Brazilian TV program since 1986, currently offers only textual versions. The overarching goal is to annotate it with morphological and syntactic characteristics, along with audio transcription annotations, emphasizing conversational dynamics.

1 Introduction

This paper outlines the preliminary efforts and chalenges in constructing an audio-transcripiton corpus within the interview genre, poised for exploration by linguists, computer scientists, historians, sociologists, political scientists, communication professionals, and researchers engaged in digital humanities.

While we currently present the initial textual versions of the corpus, our ultimate goal is to annotate it with morphological and syntactic characteristics through the Universal Dependencies methodology (De Marneffe et al., 2021). The Universal Dependencies framework is an initiative aimed at creating consistent morphosyntactic annotation for languages. The methodology encompasses the morphological characteristics annotation (POS and inflections such as gender, aspect, tense, and others) of the sentence elements as well as the relationships between them.. Additionally, we aim to incorporate an annotation related to audio-transcription, highlighting significant aspects of the conversational flow and interactions between interviewees and interviewers.

The Roda Viva corpus comprises transcribed and rewritten interviews, transformed into journalistic texts sourced from the **TV Cultura** show **Roda Viva**. Although our ultimate goal is to establish a multimodal corpus with two subcorpora (one for speech and the other for the transcription), at this moment, only the textual dataset is currently available.

The **Roda Viva** is a renowned interview show on TV Cultura, on air since 1986, and is one of the mainstays of Brazilian television. Airing on Mondays, the program hosts political figures, artists, scientists, and intellectuals, with interviews conducted by journalists or professionals from various fields. In 2007, Fapesp initiated the **Portal Roda Viva**[1] project, resulting in the **Memória Roda Viva**[2] portal providing complete transcriptions of 713 interviews, with 556,671 sentences, conducted between January 1986 and July 2009.

While the portal is a valuable resource, it has been the subject of academic study in only two works, Botin (2016) and Pacheco (2020). Despite being a rich resource, the **Memória Roda Viva** lacks formalization as a linguistically constructed corpus. Therefore, we intend to present the information available on the portal in a structured linguistic corpus and discuss transcription interventions identified throughout processing.

2 Related works

The previously mentioned works in the introduction, Botin (2016) and Pacheco (2020) are directly associated with the Memória Roda Viva project and primarily conduct theoretical linguistic analyses, rather than focusing on Natural Language Processing (NLP).

With the advancement of Automatic Speech Recognition (ASR), there has been a surge in the creation of multimodal corpora, predominantly cen-

[1] https://bv.fapesp.br/pt/auxilios/23029/projeto-portal-roda-viva/

[2] https://rodaviva.fapesp.br/

tered on audio-transcription, for NLP. Recently several datasets aimed at ASR have been introduced for Brazilian Portuguese (BP), including the CETUC dataset (Alencar and Alcaim, 2008), comprising 145 hours of references and 1,000 sentences spoken by various speakers; Common Voice Corpus 6.1 (Ardila et al., 2020), version pt_63h_2020-12-11, containing a total of 63 hours of audio from 1,120 different speakers; Multilingual LibriSpeech (MLS) dataset (Pratap et al., 2020), providing a total of 3.7 hours of audio for BP; Multilingual TEDx Corpus (Salesky et al., 2021), featuring 164 hours and 93k sentences for BP; and, more recently, CORAA ASR (Candido Junior et al., 2023), encompassing a total of 692.13 hours of audio in BP.

Regarding works predating 2020, noteworthy projects include the NURC (Norma Urbana Culta) (Silva, 1996), which have been active for over 30 years across five different Brazilian capitals (Recife, Salvador, São Paulo, Rio de Janeiro, and Porto Alegre). These projects involve the collection and provision of audio-transcriptions representing diverse manifestations of Brazilian Portuguese.

Additionally, the C-Oral-Brasil-I corpus (Raso and Mello, 2012), associated with the C-Oral-Brasil project (Raso et al., 2015), is a collection of audio transcriptions comprising 139 texts with a cumulative duration of 21.13 hours of audio.

This type of corpus is interesting both for linguistic research and for Digital Humanities studies. One inspiration for the format was the work of (Escouflaire et al., 2023), who presented 13 years of news from the Belgian French television news site.

3 Corpus Construction

The initial version of the corpus consists of all interviews available on the Memória Roda Viva Portal, in a total of 713 documents. Each document is an interview transcribed and rewritten in a journalistic format. This first set consists of 556,671 sentences, 9,432,547 tokens and 2,606,013 types, as displayed in Table 1.

Version	Sentences	Tokens	Types
V0.1	556,671	9,432,547	2,606,013
V0.2	542,716	8,996,276	2,420,553

Table 1: Corpus Textual Data

Utilizing the first interview dataset, we conducted searches for corresponding editions on the YouTube channel[3] and the show's website[4]. As a result, we identified a total of 364 interviews with an accompanying video version, as depicted in Table 1:

Interviews	Output
With video	364
Without vídeo	349
Total	713

Table 2: Roda Viva Corpus Interviews with available video

Language	Quantity
PT-BR	308
PT-BR/English	22
PT-BR/Spanish	16
PT-BR/PT-EU	8
PT-BR/French	6
PT-BR/Italian	1

Table 3: Languages distribution of the Interviews with Available Videos

Language	Duration
All languages	522h 06min 46s
PT-BR	446h 18min 49s

Table 4: Languages distribution of the Interviews with Available Videos

Subsequently, we seek to confirm which of these interviews were conducted in Brazilian Portuguese, taking into account Portuguese interviewees who reside in Brazil, such as Maria da Conceição Tavares in the maria_da_conceicao_tavares_1995 interview available on corpus. As illustrated in Table 2, 308 of 364 interviews were conducted in Brazilian Portuguese.

The 308 interviews in Brazilian Portuguese represent an amount of 466 hours of video, as highlighted in Table 3 and 4.

3.1 Current Versions

The textual corpus is available in two versions, as shown in Table 1. The first version, Version 0.1, comprises texts acquired directly from the Roda Viva portal. These texts have undergone a cleaning process, during which elements such as links

[3] https://www.youtube.com/rodaviva

[4] https://cultura.uol.com.br/programas/rodaviva/

and icons found on the original pages of each interview were eliminated. The primary goal was to meticulously preserve the textual content of the interviews, ensuring the comprehensive retention of their original form.

The second version of the corpus arose due to the need to remove the interventions made by the transcribers in the original texts. These interventions, which will be elaborated upon in the subsequent section, involved the insertion of textual information not present in the interview videos. These additions, marked as comments within square brackets ([]), were excluded in Version 0.2 of the corpus. While the disparity between version 0.2 and 0.1 is 2.5% in terms of the number of sentences and 4.6% for tokens, these interventions could introduce noise during textual processing.

Consequently, the Version 0.2 represents the transcribed texts without the introduced interventions by the transcribers.

Both versions are accessible in two formats: a compilation of CSV files (with one interview per file) and a JSON file encompassing all interviews. In CSV files there are five columns: DATA (the date of the interview in the format DD/MM/YYYY), ENTREVISTA, (the name of the interviewee), ORDEM (the order of speech in the interview), LOCUTOR (the name of the speaker) and FALA (the textualisation of the speaker's speech). These resources are available on our GitHub page[5].

4 Transcription Intervention

In the original files, each interview contains certain textual interventions, always enclosed within square brackets. These interventions result from the retextualization process undertaken during transcriptions. They manifest in various forms, with some being predictable and frequent, such as the completion of words omitted during speech:

(1) Eu acho que **[ele]** é o melhor do mundo como chargista
I think ***[he]*** *is the best cartoonist in the world*

(2) Vocês não acreditam, **[mas esse assunto]** não me preocupa agora
You don't believe it, ***[but this issue]*** *doesn't concern me now*

Or pertaining to the conversational flow, addressing the interaction among interview participants:

(3) Se você..., na eleição..., poderia fazer... **[falando junto com o Markun e concordando com ele]**
If you..., in the elections..., you could... ***[speaking together with Markun and agreeing with him]***

(4) ele fez... **[imita a pessoa respirando fundo]** Eu disse: matei o velho! **[Risos]**
he did... ***[imitates someone taking a deep breath]*** *I said: I killed the old man!* **[Laughs]**

And also occurring as abbreviations and acronyms:

(5) PIB **[produto Interno Bruto]**
PIB ***[Gross Domestic Product]***

(6) PT **[Partido dos Trabalhadores]**
PT ***[Workers Party Brazil]***

Occurrences of other interventions are less frequent and predictable. These instances typically involve explanations about a subject discussed during the speech, often assuming an encyclopedic character. The following examples have one ocurrence in the corpus:

(7) O Paulinho **[Paulinho da Viola, cantor e compositor]** gravou o quê?
Paulinho ***[Paulinho da Viola, singer and composer]*** *recorded what?*

(8) Quer dizer, saber como é que isso transformou os países da "cortina de ferro" **[expressão criada em 1946 pelo ex-primeiro-ministro britânico, Sir Winston Churchill, para designar a política de isolamento adotada pela União Soviética e seus estados-satélites após a Segunda Guerra Mundial. Foi uma expressão usada no Ocidente para designar a fronteira imaginária que dividiu a Europa em duas áreas de distintas: os países socialistas e os países capitalistas**]?
I mean, knowing how this change "Iron Curtain" countries ***[an expression created in 1946 by former British Prime Minister Sir Winston Churchill to designate the isolation policy adopted by the Soviet Union and its satellite states after The Second World War. It was an expression used in the West to designate the imaginary bor-***

[5] https://github.com/<ANONYMIZED>/Roda-Viva

> ***der that divided Europe into two distinct areas: socialist countries and capitalist countries]?***

Although these interventions constitute a minor portion of the corpus, there ara 35,663 unique occurrences of them, and they may contain pertinent information about what was said, as exemplified in (**1**) and (**2**), or about significant interactions among interviewers and interviewees, as demonstrated in (**3**) and (**4**).

As one of the upcoming steps, we aim to establish a taxonomy for these unique interventions and implement an annotation process to classify and, when necessary, differentiate them from the text.

5 Conclusion and future steps

In this initial overview of the corpus, it has become evident that the interventions conducted by the transcribers, despite constituting a smaller portion of the corpus, are highly relevant to the ultimate goal of annotation – covering both morphological and syntactic aspects as well as conversational elements. A dedicated annotation distinguishing elements such as word completion, conversational flow, abbreviations/acronyms, and topic explanations is crucial for ensuring a comprehensive and accurate version of the corpus.

Following the annotation of transcription interventions, the subsequent steps entail the automatic annotation of the corpus using the Universal Dependencies guidelines[6] – a framework for the uniform annotation of grammar, encompassing parts of speech, morphological features, and syntactic dependencies, across various human languages – through the parser being developed by the POeTiSA[7] (POrtuguese processing – Towards Syntactic Analysis and parsing) project. This annotation will undergo review and validation by linguists.

The corpus is currently undergoing a review of annotations conducted by the POS annotator Porttagger (Silva et al., 2023), which specializes in Brazilian Portuguese, developed by the previously mentioned research group.

Additionally, we aim to conduct annotations focusing on semantic-discursive aspects, specifically emphasizing translation interventions that convey irony and sarcasm (Pedro, 2018). In conjunction with morpho-syntactic annotation, we plan to identify negative triggers and their respective scope elements.

Therefore, upcoming versions of the Roda Viva corpus will incorporate the described annotation, along with the availability of audio and videos corresponding to each interview.

Acknowledgements

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #19/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

Isaac Souza de Miranda Jr. was supported by the São Paulo Research Foundation (FAPESP grant #23/01892-4).

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.

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.

Livia Maria Botin. 2016. *Ciência e tecnologia em debate: uma análise das entrevistas do programa Roda Viva, da TV Cultura*. Phd thesis, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, Brasil.

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, et al. 2023. Coraa asr: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *Language Resources and Evaluation*, 57(3):1139–1171.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal

[6] https://universaldependencies.org/

[7] https://sites.google.com/icmc.usp.br/poetisa

dependencies. *Computational linguistics*, 47(2):255–308.

Louis Escouflaire, Jérémie Bogaert, Antonin Descampe, and Cédrick Fairon. 2023. The RTBF corpus: a dataset of 750,000 belgian french news articles published between 2008 and 2021.

Priscilla Hoelz Pacheco. 2020. *A construção "acontece que" no português brasileiro contemporâneo : Uma análise baseada no uso*. Master thesis, Universidade Federal Fluminense, Niterói, Brasil.

Gabriela Wick Pedro. 2018. *ComentCorpus: identificação e pistas linguísticas para detecção de ironia no português do Brasil*. Master thesis, Universidade Federal de São Carlos, São Carlos, Brasil.

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.

Tommaso Raso and Heliana Mello. 2012. The c-oral-brasil i: Reference corpus for informal spoken brazilian portuguese. In *Computational Processing of the Portuguese Language*, pages 362–367, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tommaso Raso, Heliana Mello, Maryualê Mittmann, et al. 2015. O projeto c-oral-brasil. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 1:31–67.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech*.

Emanuel Huber da Silva, Thiago Alexandre Salgueiro Pardo, and Norton Trevisan Roman. 2023. Etiquetagem morfossintática multigênero para o português do brasil segundo o modelo" universal dependencies". *Anais*.

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