The F1 of Formula One: Applicability of Pre-trained NER Models to Brazilian TV Interview Transcripts

João Pedro Gonçalves Munhoz[1], Luiz Felipe Guidorizzi de Oliveira[2], Isabella Belchior[1], Evandro Eduardo Seron Ruiz[2] e Oto Araújo Vale[1]
[1]Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
13365-905 São Carlos, SP – Brasil,
[2]Departamento de Computação e Matemática, Universidade de São Paulo (USP)
14040-901 Ribeirão Preto, SP – Brasil
Correspondence: otovale@ufscar.br

Abstract

Recorded interviews can capture their subjects' memories, perceptions, and emotions. When conducted with notable figures, they also have the potential to serve as a resource for interdisciplinary research, impacting various branches of science. In this work, we mark the beginning of a significant project analyzing interviews from the Roda Viva program, the longest-running interview show on Brazilian television. In this initial study, we examined six memorable interviews with six Brazilian Formula One drivers to compare the performance of two named entity recognition methods: a statistical-neural method and large language models, both evaluated against manual annotations. Still, it highlighted relevant qualitative distinctions: the statistical method showed a rigid dependence on capitalisation and lexical familiarity, leading to mechanical false positives and missing non-capitalised entities, while the LLM exhibited greater linguistic sensitivity, retrieving contextual entities and being robust to transcription errors, though it still produces false positives. The LLM-based model appears more promising due to its flexibility and the potential for refinement via instructions to filter for ambiguities, favouring the automation of social network extraction in the corpus.

1 Introduction

Television interviews have been increasingly consolidated as fundamental sources for research in digital humanities, offering a rich multimodal repository that combines verbal language, body expression, intonation, and visual context, elements essential for reconstructing historical narratives and mapping social networks over time. Unlike static textual documents, these audiovisual records capture not only what was said, but also how, when, and by whom, enabling denser analyses of collective memory construction, the formation of national imaginaries, and the dynamics of power relations within specific historical contexts. In this landscape, long-form interview programs such as *Roda Viva*[1] assume particular relevance, as they constitute continuous documentary series that have recorded, over decades, the voices of leading figures in Brazilian politics, culture, science, and sports.

The current Roda Viva corpus comprises 713 transcribed interviews available on the Memória Roda Viva portal[2], compiled into machine-readable formats by de Miranda Jr et al., 2024. This corpus documents more than three decades of contemporary Brazilian history, forming a documentary repository of inestimable value for understanding the formation of our collective memory.

Ideally, comprehensive processing of this corpus would allow, among other things, the construction of an extensive named entity network—a digital map of the personal, political, and cultural connections woven across thousands of hours of public dialogue. However, before undertaking analysis at such scale, rigorous methodological validation of entity extraction and identification techniques is required.

Although transcribed interviews do not faithfully reproduce what occurred during the televised interviews, they can be considered a genre in their own right. They do not fully conform to traditional written texts, since the sequence of utterances (marked by interruptions and frequent turn-taking) differs significantly from the patterns found in news journalism or opinion pieces. On the other hand, it is evident that there are multiple levels of orality representation. While hesitations and repetitions are generally omitted to ensure readability, the flow of information retains the essential characteristic of this type of Roda Viva interview format: that of an interviewee responding alternately to multiple interviewers, thereby conferring upon the text a

[1]https://cultura.uol.com.br/programas/rodaviva/
[2]https://rodaviva.fapesp.br/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 303–311
April 13-16, 2026

dialogic and fragmented structure.

In this paper, we present a pilot study that takes as its test case a cohesive and culturally significant subset of the corpus: interviews conducted with Brazilian Formula One drivers. This selection is justified not only by the controlled yet relationally complex environment it offers for evaluating named entity recognition techniques, but also by the domain's symbolic relevance. Interviews in this subset contain dense networks of references to on-track rivals, team principals, engineers, sponsors, specialized journalists, and family members—relationships that are historically documented and amenable to systematic computational analysis.

Formula One occupies a singular place in the national imaginary: more than an elite sport, it has become a stage for projecting Brazilian identity on the global scene, with its drivers assuming roles as cultural ambassadors and, in emblematic cases, national heroes. Within this context, Ayrton Senna transcended the racetrack to become a symbol of excellence, determination, and patriotism: a contemporary myth whose trajectory, prematurely cut short, continues to shape narratives about Brazil and its place in the world. Senna's centrality as a catalyst of memories and social connections makes this subset especially fertile for investigating how public narratives construct and preserve symbolic bonds among individuals, institutions, and the nation.

This subcorpus includes archived interviews with Ayrton Senna (1986), Nelson Piquet (1994), Emerson Fittipaldi (1995), and Rubens Barrichello (1996), supplemented by our own transcriptions of interviews with Christian Fittipaldi (1995) and Lucas di Grassi (2022). By focusing on this specific domain, we aim to establish and validate a robust pipeline for Proper Name Recognition. This approach ensures that the extraction criteria remain reliable and transparent, providing a consistent framework for the future expansion of the study to the full corpus. The results discussed here not only contribute to the preservation and critical analysis of Brazilian sports memory, highlighting the legacy of its greatest hero, but also pave the methodological path for large-scale relational network construction from multimodal audiovisual sources.

The remainder of this paper is organized as follows: Section 2 reviews related works to situate our study within the current literature. Section 3 details the data and methods employed in our analysis, followed by a comprehensive presentation and discussion of our results in Section 4. The paper concludes in Section 5 with a summary of findings.

2 Related work

Automated extraction of content and relationships between those contents from interviews has been little explored, according to our literature review. Husevåg, 2019 investigates the potential of subtitles as a source for automatic indexing of TV programs through named entity recognition (NER), finding that while subtitles capture a substantial subset of salient entities, especially personal names across genres and creative works in literature programs, they alone cannot fully replicate manually created metadata. In Adriansen, 2012, the authors address one of the first studies of interviews as material of historical and social interest, explaining how researchers can utilize interviews as a tool for conducting life history research. We also found very few academic articles on enhancing historical knowledge from interviews, and the automated extraction of social relations described in these interviews remains a relatively unexplored topic, as noted by (Laato et al., 2025). Laato and his team conducted a zero-shot information-extraction study on 89,339 brief Finnish interviews with refugee families relocated after WWII. They extract social organizations and hobbies for each family member as proxies for social integration, and compare several generative models using a supervised approach to evaluate their relative strengths.

In another study, Hicke et al., 2025 has shown that researchers can expand their understanding of history and society with the help of Natural Language Processing resources and large language models. In their article, Hicke and co-workers adapted human-annotated prompts for large language models to identify and characterize portrayals of acts of God in a corpus of 88 Christian fiction novels. Similarly, (Poibeau, 2024) assessed large language models for annotating Roman and Greek mythological references in modern French literature, presented an annotation scheme, and showed how LLMs can follow it effectively despite occasional significant errors. His study includes graphically relating people, organizations, and other entities mentioned in these interviews.

Researchers recognize that the potential social relations between named individuals in the interviews involve the computational task of Named Entity

Recognition (NER). This task identifies mentions of rigid designators in free text related to predefined semantic types, such as persons, places, and organizations. A rigid designator is defined as a term that identifies the same entity across all 'possible worlds' in which that entity exists (Kripke, 1972). For instance, the name 'Aristotle' functions as a rigid designator because it refers to a specific individual regardless of the counterfactual circumstances or descriptions associated with him. Li et al., 2020 have published a valuable survey on NER.

The field of Named Entity Recognition (NER) in Portuguese has matured significantly since the inaugural HAREM evaluation contests (Santos et al., 2006; Freitas et al., 2010), progressing toward contemporary benchmarks that test the efficacy of Large Language Models (LLMs) in specialized domains. However, while we acknowledge the significant initiatives advancing the state-of-the-art—most notably the pre-trained BERTimbau transformer (Souza et al., 2020) and similar architectures (Souza et al., 2023) applied in sectors ranging from healthcare (Schneider et al., 2020) to jurisprudence (Nunes et al., 2024) – research remains predominantly focused on conventional written corpora. These genres are typically characterized by an objective or declarative tone, which contrasts sharply with the dialogic and spontaneous nature of transcribed interviews.

Consequently, the primary objective of this study is to evaluate standard off-the-shelf approaches applied to real-time, semi-spontaneous conversations, specifically television interviews. Consequently, we focus our evaluation on two distinct architectures: a dedicated neural model from an industry-standard NLP library and an open-weight Large Language Model (LLM).

3 Data & Methods

Data

The program Roda Viva is one of the longest-running interview programs on Brazilian television[1]. It has aired every Monday at 22:00 on TV Cultura since 1986. You can freely access all 27 seasons of interviews on the program's YouTube channel[2]. Researchers, students, viewers, and internet users can explore 713 transcribed interviews (de Miranda Jr et al., 2024)[3], which provide content in text form, complete with entries, references, photographs, and short videos.

In this initial project, we compare named-entity retrieval methods for extracting person mentions from Roda Viva interviews, evaluating a statistical approach, large language models, and manual annotation to see how effectively each captures the social links described in free text.

We selected interviews with six Brazilian Formula One drivers:

1. Ayrton Senna da Silva. Three-time Formula One World Champion (1988, 1990, 1991). Interviewed in 1986;

2. Nelson Piquet. Three-time Formula One World Champion, who competed in 204 races from 1978 to 1991, with 23 victories during that period. Interviewed in 1994;

3. Rubens Barrichello. Held the record for the longest uninterrupted participation in the Formula One World Championship from 1993 to 2011. Interviewed in 1995;

4. Christian Fittipaldi. Participated in 43 Formula One races between 1992 and 1994. Interviewed in 1995; and

5. Emerson Fittipaldi. Two-time Formula One World Champion (1972, 1974). Interviewed in 1995.

6. Lucas di Grassi. Participated in 18 Formula One races in 2010. 2016 FIA[4] Endurance Vice-Champion. Interviewed in 2022.

Table 1 presents descriptive statistics regarding the distribution of entities within the interviews.

Interview	Unique entities	Number of sentences	Entity density
Senna	66	208	122.59
Piquet	81	309	91.24
Barrichello	103	277	80.22
Christian	73	243	115.17
Emerson	124	269	60.15
Di Grassi	73	214	119.50

Table 1: Descriptive statistics of the dataset, showing the number of unique entities, number of sentences, and entity density.

[3] https://github.com/LeGOS-UFSCar/Roda-Viva/tree/main/Corpus/V0-2/csv

[4] Fédération Internationale de l'Automobile.

Methods

We retrieved the annotated text of the six interviews. Table 2 lists the total number of tokens in these interviews.

For all interviews, we extracted the names mentioned by the interviewee in three different ways:

1. Manual annotation. Four linguists from the Department of Letters at the Federal University of São Carlos (UFSCar) manually annotated all six interviews. Since two interviews were not included in the Projeto Memória Roda Viva[5], the linguists also performed the manual transcriptions before annotating them.

2. Neural statistical method. We used a neural transition-based model with a convolutional feature extractor (CNN) and residual connections for named entity recognition, as described in (Honnibal and Montani, 2017). We implemented this using the spaCy module and loaded the auxiliary model `pt_core_news_lg`.

3. Large language model (LLM) prompt. For this project, we adopted Ollama's `gpt-oss:20B`[6], a recent open-weight model that Ollama[7] and OpenAI[8] developed in partnership.

4 Results

Table 2 shows the number of tokens per interview.

Interview	# of tokens
Senna	15,814
Piquet	23267
Barrichello	22,863
Christian	22,688
Emerson	23,278
Di Grassi	26,050

Table 2: Number of tokens per interview.

We annotated all six interviews using the three methods described above, focusing solely on identifying personal names. We consider manual annotation the gold standard and validate the other annotations against it.

Although relatively normalized during the transcription process, these interviews, as records of oral speech, exhibit marks of orality and a degree of spontaneity characteristic of the Roda Viva program. The interviewees are seated in the center, with the interviewers arranged in a semicircle around them. In every interview, there is a moderator whose role is to facilitate communication and to ensure an impartial, balanced, and productive process. While the program's tradition confers a formal character on the interviews, the spontaneity of the dialogue means that name designations are not always uniform. Let us consider, for example, some of the personal names used in the interview with former driver Ayrton Senna: 'Ayrton Senna da Silva', his full name, mentioned only once; 'Senna', once; 'Ayrton', fifteen times; 'Ayrton Senna', seven times. In other words, there are several denominations that all refer to the same person. This variability of expression does not occur only with the interviewee, that is, the person with the longest speaking time in the interview, but also with other named individuals, such as 'Lauda' and 'Nick Lauda'[9], a three-time Formula One World Champion and, for two seasons, in 32 Grands Prix, Senna's opponent. Similar cases occur in the other four interviews and also in the recognition method based on an LLM. As a point of interest, still in the case of Ayrton Senna's interview, while the neural statistical method recognized 57 named individuals, the LLM recognized 74. The gold standard identified 66 individuals.

Although multiple denominations for the same person occur, we compared the names identified by neural statistical and LLM-based methods with the gold standard at the character level. In this approach, we considered two names identical if their characters matched.

Evaluation

To evaluate the performance of each method, we report precision, recall, and the F1-score, where precision measures the proportion of retrieved items that are relevant, recall measures the proportion of relevant items that are successfully retrieved, and the F1-score is the harmonic mean of precision and recall, providing a single summary indicator that balances both dimensions. Formally, these metrics are given by:

[5] https://rodaviva.fapesp.br/materia/207/roda_viva/sobre_o_projeto.htm

[6] https://ollama.com/library/gpt-oss

[7] https://ollama.com/

[8] https://openai.com/

[9] Sic. The entity refers to Niki Lauda; the form "Nick" appears in the source transcription.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- TP (True Positives): are the correct positive predictions;
- FP (False Positives): are the predicted positive but actually negative; and,
- FN (False Negatives): stands for the predicted negative but actually positive

Table 3 shows the TP and FP counts for both methods, the statistical neural method, and the LLM. It also shows the counts for names not found.

Interview	TP	FP	Not Found
Senna	52	5	14
Senna (LLM)	62	13	4
Piquet	62	23	19
Piquet (LLM)	68	16	13
Barrichello	75	18	28
Barrichello (LLM)	93	21	10
Christian	51	23	22
Christian (LLM)	64	33	9
Emerson	100	19	24
Emerson (LLM)	102	11	22
Di Grassi	56	28	17
Di Grassi (LLM)	64	15	9

Table 3: Counts for true positives, false positives, and names not found for the statistical neural method and the LLM.

Next, Table 5 reports the precision, recall, and F1-score of the neural statistical model for each interview. Overall, the neural model achieves acceptable recall across all interviews, indicating that it retrieves most of the names in the gold standard. However, precision varies substantially across interviews: while it is strong for Senna (0.91) and moderately high for Emerson (0.84) and Barrichello (0.81), it is lower for Piquet (0.76) and Christian (0.70), suggesting a higher rate of false positives in these cases.

The confusion matrices generated through the application of the neural statistical method, presented in Table 4, substantiate this condition. This discrepancy results in F1-scores that are reasonably balanced only for Senna (0.84) and Emerson (0.82), whereas Piquet, Barrichello, Di Grassi and Christian show considerably weaker overall performance (0.75, 0.76, 0.71 and 0.69, respectively).

Table 4: Confusion Matrices under the statistical neural method

		Predicted	
		Positive	**Negative**
Actual	**Pos**	62 (TP)	19 (FN)
	Neg	23 (FP)	0

(a) Performance for the Piquet interview.

		Predicted	
		Positive	**Negative**
Actual	**Pos**	51 (TP)	22 (FN)
	Neg	23 (FP)	0

(b) Performance for the Christian interview.

In the Senna interview, the model achieves an F1-score of 0.84, the highest among all cases for this method. The Table 5, therefore, reveals that the model is not uniformly robust across interviews and may be sensitive to interview-specific characteristics, such as lexical variation, discourse structure, or annotation idiosyncrasies.

Interview	Precision	Recall	F1-score
Senna	0.91	0.79	0.84
Piquet	0.73	0.76	0.75
Barrichello	0.81	0.73	0.76
Christian	0.69	0.70	0.69
Emerson	0.84	0.81	0.82
Di Grassi	0.67	0.77	0.71

Table 5: Evaluation metrics for the neural statistical model.

Similarly, Table 7 presents the precision, recall, and F1-score of the LLM-based method for each interview. Compared to the neural statistical model, the LLM achieves systematically higher recall, particularly for the Senna, Barrichello, and Christian interviews (all with recall greater than or equal to

0.88), indicating that it retrieves a larger proportion of the names present in the gold standard. However, this gain in recall may come at the expense of precision, which remains very low for Christian (0.66), revealing a substantial number of false positives and suggesting a tendency to over-generate named entities. The confusion matrices obtained from the implementation of the LLM method, as illustrated in Table 6, provide compelling evidence for this condition. As a result, the F1-scores for the Christian interview (0.75) remain modest, and all the others show clearly reasonable overall performance. We also see that, in the case of Senna, both recall and precision are high. Thus, while the LLM model is effective in not "missing" names, it lacks consistent reliability across interviews and appears particularly prone to spurious name recognition in a single case.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	64 (TP)	9 (FN)
	Negative	33 (FP)	TN

Table 6: Confusion Matrix for Christian under the LLM method.

Interview	**Precision**	**Recall**	**F1-score**
Senna	0.83	0.94	0.88
Piquet	0.81	0.84	0.82
Barrichello	0.82	0.90	0.86
Christian	0.66	0.88	0.75
Emerson	0.90	0.82	0.86
Di Grassi	0.81	0.88	0.84

Table 7: Evaluation metrics for the LLM model.

A rough comparison of these two approaches highlights differences among the three metrics. In Table 8, we display the percentile differences between the valuation metrics for the LLM model and the statistical neural model.

With the exception of precision in the Senna and Christian interviews, the Ollama LLM outperformed the statistical neural model.

Interview	**Precision**	**Recall**	**F1-score**
Senna	-0.08	0.15	0.04
Piquet	0.08	0.08	0.07
Barrichello	0.01	0.17	0.10
Christian	-0.03	0.18	0.06
Emerson	0.06	0.01	0.04
Di Grassi	0.14	0.11	0.13

Table 8: Estimating the difference in percentile between the valuation metrics for the LLM model and the statistical neural model.

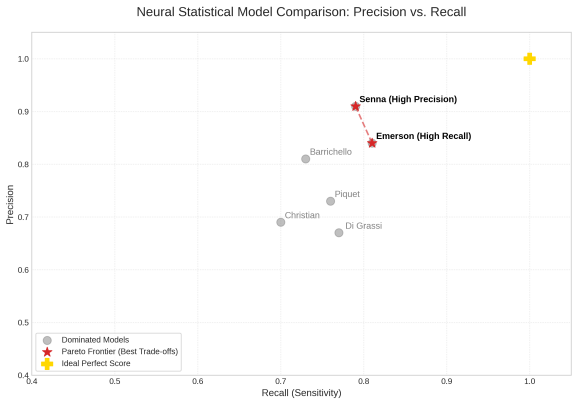

Figure 1: Comparison of precision X recall for the neural statistical model.

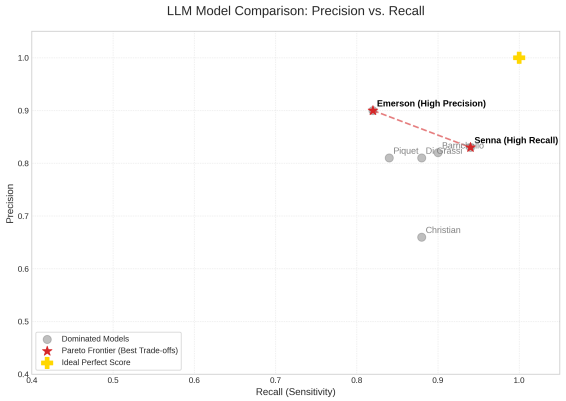

Figure 2: Comparison of precision X recall for the LLM model.

Figures 1 and 2 above illustrate the precision-recall performance for both models. Additionally, we have drawn a Pareto line (or Pareto Frontier) to delineate the boundary of optimal performance. This frontier connects the set of 'non-dominated' configurations, implying that no other option offers a superior combination of both metrics. Consequently, points lying on this line represent the most efficient trade-offs available.

5 Conclusion

This study evaluated two distinct approaches for retrieving named entities from the Roda Viva interview archive: a neural statistical method and a large language model. Quantitatively, the difference between the two methods was not statistically significant as initially expected, with p-values consistently exceeding the 0.05 threshold across all six interviews analyzed. However, a qualitative analysis of errors and 'false positives' reveals a fundamental divergence in how each model interprets the concept of a named entity in Brazilian Portuguese.

The neural statistical method demonstrated a rigid reliance on lexical familiarity and orthographic features, specifically capitalization. While this aligns with standard Portuguese grammar, where proper nouns use title casing, the neural model struggled to differentiate between named entities and simple sentence initiators, incorrectly tagging phrases such as *'É. Exato'* and *'É. Minha'* solely due to their casing. Furthermore, the neural statistical model appeared to treat lexical anomalies as probable proper nouns. Tokens such as *'eh'* and *'Ã'* (likely unintelligible to the model lexically) were frequently tagged as entities, suggesting that the model categorizes unknown or out-of-vocabulary terms as proper names by default. Conversely, the model failed to identify valid entities that lacked standard capitalization, resulting in lower recall in complex scenarios than the generative approach.

In contrast, the LLM demonstrated greater productivity and linguistic sensitivity, showing less sensitivity to capitalization and a more semantic focus. It successfully retrieved contextually correct entities that human annotators missed, such as *'a mãe do Rubinho'* (Rubinho's mother) and *'tio Emerson'* (Uncle Emerson). Notably, the LLM demonstrated robustness to transcription errors: it correctly identified *'Ayrotn'* (a typo introduced by transcribers from the previous project) as a named entity, whereas the neural statistical method failed to detect it. While it remains to be seen if the LLM can successfully link this orthographically deviant form to the specific 'Ayrton Senna' entity in a downstream resolution task, this detection capability highlights the model's semantic focus over strict orthographic matching. This is highly valuable for Digital Humanities, where transcriptions can suffer from archival inconsistencies or transcription errors such as 'Ayrotn'. The capacity of LLMs to prioritize semantic context over rigid character matching makes them a more reliable resource for the unsupervised, *en masse* processing of unlabeled historical text, ensuring that actors within the archive remain visible even when the digital record is imperfect.

Interestingly, both models showed similar limitations in entity boundary assignment in possessive constructions. In instances such as *'meu filho Luca'* (my son Luca) and *'minha filha Joana'* (my daughter Joana), both the statistical method and the LLM extracted only the proper names (*'Luca'*, *'Joana'*) rather than the complete descriptive noun phrase. This suggests that, without a defined framework explicitly instructing the methods to identify the longest possible entity span, both approaches tend to default to the specific proper noun rather than the relational context.

While the LLM's sensitivity led to specific types of false positives that differed significantly from the mechanical errors of the statistical model, these results offer a unique benefit for downstream processing. The LLM frequently annotated highly deictic expressions such as *'seu pai'* (your father), *'tua irmã'* (your sister), and *'meu pai'* (my father), as well as personal pronouns. Although these terms semantically refer to persons, they were excluded from the manual gold standard because effectively incorporating them would require an additional layer of annotation focused on resolving relations between entities, i.e. coreference resolution (Liu et al., 2023).

However, rather than viewing these as simple errors, we argue that this sensitivity is a methodological advantage. In a Digital Humanities context, capturing these deictic markers is a crucial first step for entity linking and social graph extraction. If a future model is capable of correctly performing entity linking, these "false positives" become high-value nodes that link individuals through kinship and social proximity, providing a much denser map of the Roda Viva archive than proper nouns alone. Without this initial capture, such deep relational information could remain invisible to unsupervised, *en masse* processing.

It is important to note that these localized, personal references may offer diminishing returns for the project's purpose. Because these entities are often unique to a single interview's narrative, they risk remaining as isolated nodes. Unlike public figures who appear across decades of the Roda

Viva corpus, these specific relations may not contribute to the 'global scheme' of the social network, ultimately offering limited value for large-scale, cross-interview relational mapping.

Ultimately, while both methods achieved comparable F1-scores, the LLM shows greater promise for future iterations of this project. Its false positives are linguistically grounded rather than orthographically accidental, making them methodologically more manageable. The generative nature of the model allows for the implementation of improved system instructions (such as negative constraints to ignore pronouns or a stricter reference framework) to filter out these ambiguities. Therefore, despite current statistical parity, the LLM offers a more flexible and robust approach to automating the extraction of social networks from the Roda Viva corpus.

This study validates the Roda Viva corpus not only as an audiovisual archive but also as a powerful textual resource for Digital Humanities, essential for the reconstruction of historical narratives and the dynamic mapping of social networks. By demonstrating that large language models (LLMs) offer the necessary semantic flexibility to process the spontaneity of televised speech and withstand transcription errors, this study overcomes the initial methodological barrier to the large-scale processing of this corpus. Thus, the findings presented here pave the way for transforming thousands of hours of public discourse into a structured digital map of the political, cultural, and personal connections that weave the Brazilian collective memory, thereby fulfilling the purpose of rendering visible the power dynamics and national imaginaries preserved across these decades of interviews.

Acknowledgement

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 #2019/07665-4) and by the IBM Corporation.

References

Hanne Kirstine Adriansen. 2012. Timeline interviews: A tool for conducting life history research. *Qualitative Studies*, 3(1):40–55.

Isaac Souza de Miranda Jr, Gabriela Wick-Pedro, Cláudia Dias de Barros, and Oto Vale. 2024. Roda Viva boundaries: an overview of an audio-transcription corpus. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese – Vol. 2*, pages 165–169, Santiago de Compostela, Galicia, Spain. Association for Computational Linguistics.

Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. 2010. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rebecca M. M. Hicke, Brian W. Haggard, Mia Ferrante, Rayhan Khanna, and David Mimno. 2025. Are You There God? Lightweight Narrative Annotation of Christian Fiction with LMs. *arXiv preprint arXiv:2507.19756*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Anne-Stine Ruud Husevåg. 2019. From subtitles to substantial metadata: examining characteristics of named entities and their role in indexing. *International Journal on Digital Libraries*, 20(3):241–251.

Saul A. Kripke. 1972. Naming and necessity: Lectures given to the Princeton University philosophy colloquium. In *Semantics of Natural Language*, pages 253–355. Springer.

Joonatan Laato, Jenna Kanerva, John Loehr, Virpi Lummaa, and Filip Ginter. 2025. Extracting Social Connections from Finnish Karelian Refugee Interviews Using LLMs. *arXiv preprint arXiv:2502.13566*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Rafael Oleques Nunes, Dennis Giovani Balreira, André Suslik Spritzer, and Carla Maria Dal Sasso Freitas. 2024. A named entity recognition approach for Portuguese legislative texts using self-learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 290–300.

Thierry Poibeau. 2024. Annotating References to Mythological Entities in French Literature. *arXiv preprint arXiv:2412.18270*.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. HAREM: An advanced NER evaluation contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Fábio Capuano de Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Fábio Capuano de Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2023. BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, 149:110901.