Text segmentation with mixed data: evaluating symbolic and neural approaches

Amanda Sammer do N. Brandão[1], Felipe Bomfim Sanches[1], Alfredo Sena[1], Laila Mota [1], Daniela Barreiro Claro[1] Rerisson Cavalcante[2]

[1]FORMAS - Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
{amandabrandao,felipe.sanches,alfredosena,laila.pereira,dclaro}@ufba.br
[2]FORMAS - Research Center on Data and Natural Language
Institute of Letters – Federal University of Bahia (UFBA)
Salvador – Bahia – Brazil
rerisson.cavalcante@ufba.br

Abstract. *Text segmentation has been widely adopted and used to extract relevant information from documents. This information is distributed across various types of structures, including heterogeneous data. The extraction of such data has recently been extensively leveraged by Large Language Models (LLMs). However, the training of these models does not inherently capture the particular characteristics of mixed data, such as lexical and phonetic information. As a result, automatically assessing the correctness of transcriptions using LLMs becomes a challenging task. In this study, text segmentation and information extraction were parallelized into two main approaches in order to compare a symbolic approach with a neural one. The results highlight the advantages and disadvantages of both approaches when dealing with mixed data types.*

Resumo. *A segmentação de texto vem sendo amplamente difundida e utilizada para extrair informação relevante de documentos. Essas informações estão distribuídas em diversos tipos de estruturas, que incluem dados heterogêneos. A extração desses dados tem sido ultimamente, utilizada por LLMs (Large Language Models). Porém, os treinamentos dos modelos de linguagem não detém características peculiares de dados mistos, tais como dados lexicais e dados fonéticos. Assim, torna-se um desafio analisar a corretude das transcrições de maneira automatizada por LLMs. Neste estudo, a segmentação textual e extração de informação foi paralelizada em duas principais abordagens com o intuito de avaliar a comparação de uma abordagem simbólica e uma abordagem neural. Os resultados demonstram as vantagens e desvantagens em ambas as abordagens para os tipos de dados mistos.*

1. Introduction

Text segmentation and information extraction are fundamental Natural Language Processing (NLP) tasks for transforming unstructured or semi-structured documents into computationally manipulable representations [Ghinassi et al. 2024, Lu et al. 2022]. In this study, segmentation refers specifically to identifying question blocks, speech turns, and their corresponding textual content in a Phonetic-Phonological Questionnaire (QFF). Such documents combine heterogeneous linguistic data, including graphematic text, phonetic transcription, numerical question identifiers, informant responses,

interviewer interventions, and speech-turn markers. Studies involving Brazilian Portuguese corpora highlight the importance of careful organization and manual validation [Candido Junior et al. 2023], while [Queiroz et al. 2023] demonstrate the relevance of explicit extraction rules, manual annotation, and expert validation in the construction of reliable linguistic resources. Phonetic transcriptions also require the preservation of symbols from the International Phonetic Alphabet (IPA) [International Phonetic Association 1999].

Recent advances have expanded the use of Large Language Models (LLMs) in segmentation and information extraction tasks [Krassovitskiy et al. 2025, Cabral et al. 2024]. However, the task investigated here requires more than semantic equivalence: question identifiers, speaker attribution, speech-turn boundaries, textual content, and phonetic symbols must be preserved with minimal structural modification. Because LLM outputs are probabilistic, plausible responses may not exactly reproduce the original structure, particularly regarding instruction following, schema adherence, and standardized output formats [Tang et al. 2024].

Our study investigates how to segment and extract information from a single QFF document while preserving its structural and symbolic layers. We hypothesize that, in documents with recurring and well-defined patterns, a symbolic rule-based approach may provide greater structural control, effectiveness, and computational efficiency than the evaluated LLM-based configuration. This hypothesis is supported by the relevance of rule-based methods for documents with predictable patterns [Li et al. 2008] and by the need for domain-specific comparisons between traditional and neural segmentation approaches [Jegan and Henrich 2025].

To investigate this hypothesis, we compare a symbolic method based on explicit rules, regular expressions, and document heuristics with an LLM-based method implemented through prompt engineering. The evaluation considers precision, recall, F1-score, processing time, structural preservation, and fidelity to the original content. The main contributions are: (i) a symbolic pipeline for segmenting question blocks and speech turns; (ii) a controlled comparison between symbolic and LLM-based extraction; and (iii) an analysis of the errors and structural variations produced by the evaluated approaches.

The remainder of this article is organized as follows. Section 2 presents the related work, Section 3 describes the methodology, Section 4 presents the experiments, Section 5 reports the results, and Section 6 discusses the main findings. Section 7 presents the final considerations and future works.

2. Related Work

Text segmentation and Information Extraction (IE) support the Transformation of unstructured or semi-structured documents into representations suitable for curation, indexing, linguistic analysis, and database construction [Ghinassi et al. 2024]. LLM-based methods have also been applied to structured extraction tasks; for example, [Gazzola et al. 2025] evaluate an LLM for product attribute value extraction against a reference dataset.

However, LLM-based approaches may present difficulties in tasks requiring strict schema adherence and exact preservation of the original structure, including inconsistent responses, output-format variation, and incomplete compliance with predefined schemas

[Tang et al. 2024]. Such limitations are especially relevant when identifiers, speaker attribution, speech-turn boundaries, textual content, and phonetic symbols must be preserved without omission or reorganization.

Symbolic methods based on explicit rules, structural patterns, and regular expressions remain relevant for documents with recurring organization [Li et al. 2008]. Domain knowledge can also support segmentation and normalization [Godoi et al. 2025]. In Portuguese IE research, [Queiroz et al. 2023] describe the transformation of automatically generated extractions into a manually validated corpus, reinforcing the importance of annotation criteria, expert revision, and reliable reference data.

Comparisons between traditional and LLM-based methods emphasize the need for domain-specific evaluation [Jegan and Henrich 2025], while [Xu et al. 2024] discuss the capabilities and limitations of generative approaches to IE. Nevertheless, limited evidence exists on the comparative performance of symbolic and LLM-based methods in linguistic questionnaires that combine question identifiers, graphematic text, phonetic transcriptions, and speech-turn markers under strict structural and symbolic fidelity requirements. Our study addresses this gap by evaluating both approaches on the same Phonetic-Phonological Questionnaire, considering precision, recall, F1-score, processing time, structural preservation, and fidelity to the extracted content.

The next section describes our methodology, including the corpus, preprocessing procedures, extraction strategies, and evaluation protocol.

3. Methodology

Our methodology is experimental and comparative, focusing on the automation of text segmentation and information extraction in a transcribed questionnaire from ALiB Project. In this work, the data are treated as heterogeneous because they bring together graphematic text, phonetic transcriptions, question markings, responses, and speech turns within the same document structure. The methodological objective is to evaluate how different approaches deal with this structure, considering the preservation of linguistic information and the organization of data in a structured format.

The process was divided into stages. Initially, the corpus was analyzed in order to identify the organization of the document, the question markers, the speech turns, and the annotation layers present in the transcription. Next, the document underwent a preprocessing stage aimed at standardizing the file format, preserving the textual content of the phonetic segments, and preparing the data for automatic processing.

Two approaches were implemented and compared. The symbolic approach uses explicit rules, regular expressions, and document heuristics to identify questions, responses, speech turns, and phonetic segments. The LLM-based neural approach uses prompt engineering, with instructions aimed at segmenting and extracting the same information. In both cases, the outputs were organized into structured files, such as CSV, containing fields such as question identification, speech turn, and extracted textual content. Figure 1 presents an overview of the two approaches evaluated.

The comparative evaluation considered effectiveness metrics, such as precision, recall, and F1-score, as well as aspects of computational efficiency, such as processing time and resource usage. The outputs produced by both approaches were compared

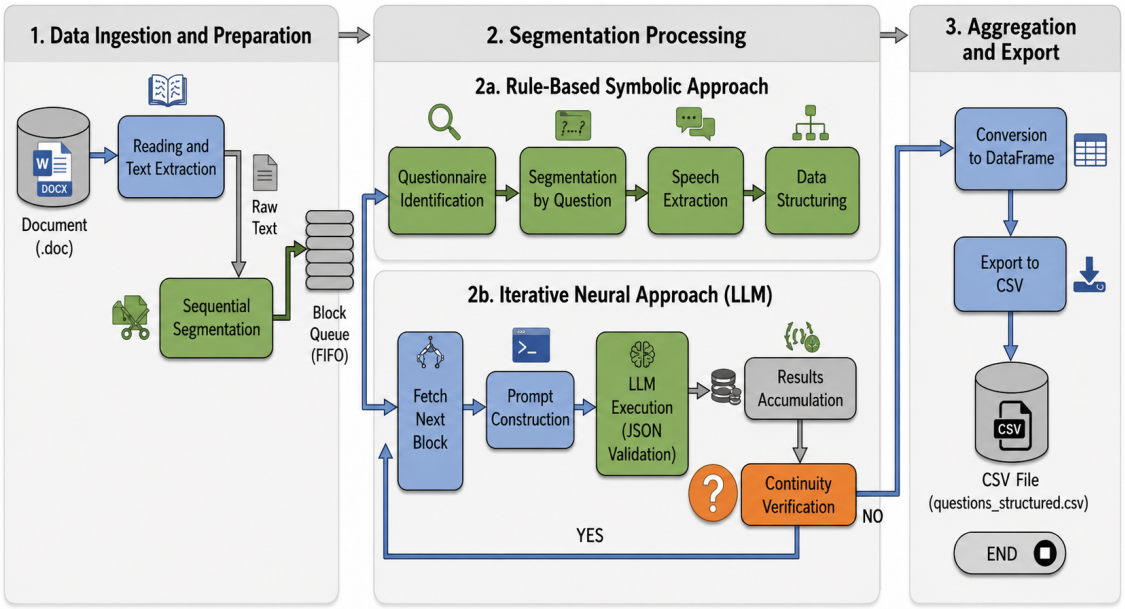

Figure 1. Overview of the symbolic and LLM-based neural approaches.

against a manually constructed gold standard derived from the original questionnaire transcription. This comparison made it possible to evaluate the ability of each method to preserve the document structure, the fields of interest, and the extracted phonetic segments.

3.1. Corpus and Preprocessing

The corpus used in our study consists of a transcribed Phonetic-Phonological Questionnaire (QFF) from Atlas Linguístico do Brasil Project (ALiB) [Cardoso et al. 2014]. The transcription brings together graphematic and phonetic information, as well as question, response, and speech-turn marks. Phonetic forms are represented by symbols from the International Phonetic Alphabet (IPA) [International Phonetic Association 1999].

The initial stage consisted of analyzing the structure of the documents and questionnaires with the objective of understanding the organization of the questions, responses, speech turns, and annotation layers. This process made it possible to define the units to be segmented and the information fields to be extracted, while maintaining alignment with the linguistic characteristics of the corpus.

The handling of phonetic transcriptions required specific care, since the correct visualization of IPA symbols depends on the compatibility among font, encoding, and file format. During data preprocessing, the documents were standardized into a textual format compatible with automatic processing, reducing inconsistencies and textually preserving the phonetic segments.

This task was not necessarily related to loss of information during textual extraction, but to the absence or non-automatic application of appropriate fonts for visualizing IPA symbols. To enable inspection of the results, it was necessary to install the SIL-DoulosIPA and Doulos SIL fonts. Even so, the software used to view the CSV files did not automatically apply these fonts to the phonetic excerpts.

For this reason, an additional manual verification stage was necessary. After the

CSV files were generated, the phonetic segments delimited by brackets were identified and the appropriate font to allow the correct visualization of IPA symbols was applied manually. This stage was validated by a linguist and proved essential for distinguishing real extraction errors from merely visual problems.

3.2. Symbolic Method

The symbolic method performs text segmentation and information extraction deterministically, based on explicit rules defined by regular expressions.

The heuristics used were defined based on the document patterns observed in the questionnaires. These rules make it possible to segment the document into questions and then identify the speech turns associated with each of them.

A) Heuristic 1: Numerical identifiers.

The first heuristic consists of identifying the numerical markers of the questions, represented in the format "(001)", "(002)", "(003)", and so on. These identifiers are used to delimit the beginning of each question and separate the document into structured blocks.

B) Heuristic 2: Speech-turn markers.

The second heuristic consists of identifying speech-turn markers, such as `INF`, `INQ`, `CIR`, and `CIRC`. These markers indicate alternation between the participants in the interaction and make it possible to associate each textual excerpt with its respective turn. When a line does not present a new marker, its content is incorporated into the previous turn, preserving speech segments distributed over more than one line.

3.3. Neural Method

Unlike the symbolic method, this second approach uses a neural method to automatically identify patterns in the transcribed texts. From this, this method categorizes the questions and answers, including the phonemes of the International Phonetic Alphabet (IPA).

The method is based on a pipeline in which the raw text from the file is extracted and divided into smaller context windows, which are used to compose the prompt, where the model receives explicit instructions on how to recognize the structure of the document and what output format it should generate. Instead of relying on previously defined rules, the model infers the patterns directly from the textual context provided.

Different context windows were tested with the objective of evaluating the impact of the size of the excerpt sent to the model both on the precision of the results and on the processing time. This variation made it possible to observe the balance between performance and computational cost.

3.4. Gold Standard Construction and Validation

To evaluate the automated extraction methods, a reference gold standard was manually constructed based on the target Phonetic-Phonological Questionnaire (QFF). To ensure high reliability and address the complexity of linguistic data, the construction and validation process involved three researchers and was carried out in three distinct phases: primary extraction, peer review, and expert consensus.

In the first phase, a primary annotator manually processed the assessment document, systematically extracting speech turns, question identifiers, and responses containing phonetic transcriptions into predefined structured fields. Subsequently, a second researcher reviewed the generated dataset to identify potential omissions or formatting inconsistencies, particularly regarding the preservation of IPA phonetic symbols.

Finally, the validation phase concluded with a consensus meeting involving a linguist. As the task required structural and symbolic fidelity, a consensus-based annotation methodology was adopted rather than parallel blind annotation. All structural alignments and data representations were subject to collective deliberation.

4. Experiments

The experiments were conducted with the objective of comparing the symbolic and neural approaches in text segmentation and extraction of speech-turn extraction from a transcribed QFF document from the ALiB Project. As a case study, the Phonetic-Phonological Questionnaire (QFF), referring to the Salvador 093-02 inquiry, was selected because it contains graphematic and phonetic transcriptions relevant to the proposed evaluation. The comparison considers effectiveness metrics, such as accuracy, precision, recall, and F1-score, as well as aspects of computational efficiency, such as processing time and resource usage. For validation of the results, the outputs produced by the methods were compared with a Gold Standard manually constructed from the original transcriptions.

The evaluation was guided by the following research question:

Q1: Which approach presents the best relationship between effectiveness and computational efficiency in text segmentation and information extraction in a transcribed questionnaire?

4.1. Evaluation of the Segmentation Strategy

In the symbolic approach, the experiments were conducted by applying the heuristics described in Subsection 3.2.1 to the document selected for evaluation. Segmentation was performed deterministically, using numerical identifiers to delimit the questions and speech-turn markers to associate each textual excerpt with its respective participant.

The generated output was structured in CSV format and manually compared with the original document in order to verify the fidelity of the segmentation, the integrity of the extracted content, and the textual preservation of the IPA phonetic transcriptions.

During the initial experiments with the LLM-based neural approach, two strategies for sending the data to the model were compared: processing the complete document and segmentation by context windows. Sending the entire document resulted in a longer response time and more inconsistencies in the extraction, indicating the need to control the volume of information provided in each request.

For this purpose, segmentation by context windows was adopted, in which the file is divided into blocks of fifteen lines of raw text, and this context is injected into the prompt sent to the local model. Since the questionnaire contains 159 questions, the impact of the size of the batches sent to the model was also evaluated. Batches with 10, 15, 25, 50, and 100 context windows were tested, considering processing time, incomplete

or inconsistent responses, and execution stability. Batches greater than 15 questions increased inconsistencies and response variability, while smaller batches increased the total execution time due to the larger number of requests. Thus, the configuration with 15 windows per batch was adopted because it presented the best balance between efficiency and extraction quality.

In order to maintain the privacy of the raw data, the tests using the LLM were performed locally without any exposure to the internet environment through Ollama v0.18.2 [Ollama 2024] with DeepSeek R1 model (8B parameters) [Ollama Library 2024] for identifying, extracting, and categorizing the questions and answers in the transcription documents.

4.2. Experimental Setup

The experiments were carried out on a document from the Phonetic-Phonological Questionnaire (QFF) of the ALiB Project, used as a case study because it contains graphematic and phonetic transcriptions relevant to the segmentation task. The original document, made available in `.doc` format, has 15 pages and contains 159 questions.

After the segmentation of the speech turns, the extractions resulted in CSV files with 211 rows, distributed across three main columns: question identification, speech turn, and textual content. The symbolic and LLM-based neural approaches were applied to the same document, allowing the comparison of the outputs produced under equivalent conditions.

For validation purposes, the outputs generated by the approaches were manually compared with the original document. The comparison aimed to verify the fidelity of the segmentation in relation to the original document, considering the integrity of the textual content and the textual preservation of the IPA phonetic transcriptions.

5. Results

The evaluation was carried out using a document from the Phonetic-Phonological Questionnaire (QFF) from the ALiB Project, composed of 15 pages and 159 questions. After the segmentation of the speech turns, the structured outputs resulted in CSV files with 211 rows, containing the question identification, the speech turn, and the extracted textual content. The outputs of the symbolic and LLM-based neural approaches were compared with a *gold standard* manually constructed from the original transcriptions.

Table 1 presents the execution times observed for the manual process, for the symbolic approach, and for the LLM-based neural approach. In the case of the LLM-based approach, the configuration of batches with 15 questions per request was considered.

Table 1. Execution time observed in the evaluated processes.

Process	Evaluated document	Execution time
Manual	1 QFF	2 h 38 min
Symbolic	1 QFF	0.3 s
LLM-based neural	1 QFF	44 min 13 s

Table 2 presents the quantitative results obtained from the comparison of the automatic outputs with the *gold standard*. The metrics of precision, recall, and F1-score were

considered, in addition to the values of true positives (TP), false positives (FP), and false negatives (FN).

Table 2. Comparison of metrics between the symbolic approach and the LLM-based neural approach.

Metric	Symbolic	LLM-based neural
Precision	98.10%	40.38%
Recall	95.81%	39.07%
F1-score	96.94%	39.72%
TP	206	84
FP	4	124
FN	9	131

The values presented in Table 2 show that the symbolic approach achieved greater correspondence with the *gold standard*. This approach reached an F1-score of 96.94%, while the LLM-based neural approach obtained an F1-score of 39.72%. The difference is also observed in the FP and FN values, indicating that the symbolic approach produced fewer incorrect extractions and omissions in relation to the reference material.

Regarding execution time, Table 1 indicates that the symbolic approach presented lower computational cost. While the manual process required 2 hours and 38 minutes, the symbolic approach processed the document in 0.3 seconds. The LLM-based neural approach, in turn, required 44 minutes and 13 seconds in the evaluated configuration.

The error analysis indicated that the LLM-based neural approach presented greater variability in the organization of the output. In some cases, the model altered the expected structure, such as by dividing the same speech segment into more lines than necessary, impairing the exact correspondence with the original document. The symbolic approach, on the other hand, presented greater structural stability, as it operates based on explicit rules applied to the observed document patterns.

Figure 2 presents examples of the behaviors observed in the outputs of the evaluated approaches. The examples were selected from the manual comparison between the original document, the output of the symbolic approach, and the output of the LLM-based neural approach. The field names *número*, *autor*, and *texto* were preserved in Portuguese because they correspond to the original fields produced by the information extraction process. In this structure, *número* indicates the question number, *autor* identifies the speaker or source, and *texto* contains the extracted textual content.

The qualitative analysis showed that the symbolic approach maintained greater adherence to the original structure, especially in the delimitation of speech turns and in the textual preservation of the phonetic segments. The LLM-based neural approach, although it preserved part of the general content, demonstrated greater variation in the organization of the output, including cases of improper division of speech segments and alteration of the expected structure. These behaviors help explain the differences observed in the quantitative metrics.

Document type	numero	autor	texto
Original Document	1	INF.	- Eh, barraco, casas ['kazɐʃ], apartamentos.
Symbolic	1	INF	- Eh, barraco, casas ['kazɐʃ], apartamentos.
Neural	1	INQ	INF. - Eh, barraco, casas ['kazɐʃ], apartamentos.

Figure 2. Qualitative examples comparing the original document and the outputs of the evaluated approaches.

6. Discussion

Our results, composed of a document from the Phonetic-Phonological Questionnaire (QFF) with 159 questions and 211 rows after the segmentation of speech turns, indicate significant differences between our approaches. Our symbolic approach achieved an F1-score of 96.94%, with 206 true positives, 4 false positives, and 9 false negatives. In comparison, the LLM-based neural approach obtained an F1-score of 39.72%, with 84 true positives, 124 false positives, and 131 false negatives. These values show greater correspondence of the symbolic approach with the *gold standard*, especially due to the lower occurrence of incorrect extractions and omissions.

This difference can be explained by the recurring structure of the analyzed transcribed questionnaires. The presence of numerical identifiers and explicit speech-turn markers favored the application of symbolic heuristics, allowing greater structural control in the segmentation and extraction process. Thus, in documents with well-defined patterns, the predictability of the format contributed directly to the precision, reproducibility, and stability of the symbolic approach.

The LLM-based neural approach presented greater output variability and lower correspondence with the original document. Since this task requires preservation of identifiers, speaker attribution, speech-turn boundaries, textual content, and phonetic symbols, semantic equivalence alone is insufficient. In the evaluated DeepSeek-R1 8B configuration, textual reorganization and improper speech-segment splitting were therefore considered extraction errors [Tang et al. 2024].

Regarding computational efficiency, the difference was also significant. While manual extraction required 2 hours and 38 minutes, the symbolic approach processed the document in 0.3 seconds. The LLM-based neural approach, in the evaluated configuration, required 44 minutes and 13 seconds. These results indicate that, for the analyzed document, the symbolic approach presented the best relationship between effectiveness and computational cost.

These findings reinforce that the choice of extraction strategy should consider the document structure, the required level of fidelity, and the available computational resources. For the analyzed QFF document, the symbolic approach was more suitable because its recurring numerical identifiers and speech-turn markers favored deterministic processing. Although LLM-based methods may be useful when greater contextual inference is required, the evaluated DeepSeek-R1 8B configuration showed lower stability in

a task that required strict structural and symbolic fidelity.

7. Final Considerations and Future Work

This study investigated the feasibility of automating segmentation and information extraction in a transcribed QFF document from the ALiB Project. The comparison between the symbolic approach and the LLM-based neural approach indicated that, in highly structured documents containing phonetic transcriptions, the symbolic approach presents a better balance between precision, computational efficiency, and reproducibility. These results suggest that symbolic methods remain relevant in scenarios in which document patterns are recurrent and structural fidelity is a determining aspect for extraction quality.

As future work, we intend to: (i) Expand the experimental corpus to include multiple ALiB questionnaires from different regions, informants, and document formats, in order to assess the external validity and robustness of the evaluated approaches and support the population of the ALiBWeb relational database; (ii) Include other LLMs and model configurations to assess whether the observed behavior is particular from DeepSeek-R1 8B; (iii) Improve the linguistic analysis based on the structured data, including morphological and dialectological studies; and (iv) Investigate the application of vector representations and phonetic embeddings for more advanced comparative analyses.

Limitations

This study focused primarily on structural and phonetic aspects of the documents, without exploring deeper semantic analyses. Our evaluation was conducted on a single QFF document from the ALiB Project, which limits the generalizability and external validity of the results and does not fully capture the variability of the complete ALiB dataset, including questionnaires from different regions, involving different informants, and available in different document formats. Although the neural approach based on Large Language Models (LLMs) was explored, in this study only one model, DeepSeek-R1 8B, was evaluated. Therefore, the results should not be generalized to LLMs in general, since different models, prompts, and configurations may produce different outcomes. This variability may affect reproducibility and requires careful validation, particularly in structured linguistic data extraction tasks. In addition, the symbolic approach relies on predefined rules and patterns, which may require manual adjustments when applied to documents with different structures or formatting inconsistencies.

Ethics Statement

This work is based on the processing of linguistic data derived from a transcribed QFF document from the ALiB Project. The data contain sensitive personal information that was properly anonymized. All materials were handled in accordance with ethical research standards, with careful attention to data integrity, transparency, reproducibility, and the prevention of harm to individuals or groups. The contact with Internet was prevent during LLM's experiments.

Use of Generative AI Tools

ChatGPT, Gemini, and Copilot were used as auxiliary tools for grammatical revision, text improvement, organization of ideas, English translation, the generation and refinement of visual materials, exploratory searches, the organization of academic references,

and support in the development and revision of Python scripts. All AI-assisted outputs were carefully reviewed, edited, tested, and validated by the authors, who remain fully responsible for the content and conclusions of the final manuscript.

Acknowledgments

This work is supported by INCT TILD-IAR (grant 408490/2024-1), FAPESB CCE022/2023 and FAPESB BOL0129/2025.

References

Cabral, B., Claro, D., and Souza, M. (2024). Exploring open information extraction for Portuguese using large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 127–136, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Candido Junior, A., Casanova, E., Soares, A. d. S., Oliveira, F. S. d., Oliveira, L., Fernandes Junior, R. C., Silva, D. P. P. d., Fayet, F. G., Carlotto, B. B., Gris, L. R. S., and Aluísio, S. M. (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.

Cardoso, S. A. M. d. S., Mota, J. A., Aguilera, V. d. A., Aragão, M. d. S. S. d., Isquerdo, A. N., Razky, A., Margotti, F. W., and Altenhofen, C. V. (2014). *Atlas Linguístico do Brasil*, volume 1. EDUEL, Londrina.

Gazzola, M., Souto, H. G., Silva, S., Peixoto, J. S., Siqueira, F., Morais, A. L. P. d., and Gomes, C. (2025). AI-PAVE-Br: Leveraging large language models for enhanced product attribute value extraction through a golden set approach. In *Proceedings of the Symposium in Information and Human Language Technology (STIL)*.

Ghinassi, I., Wang, L., Newell, C., and Purver, M. (2024). Recent trends in linear text segmentation: A survey. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3084–3095, Miami, Florida, USA. Association for Computational Linguistics.

Godoi, G. A. d., Rivolli, A., Freire, D. L., Pereira, F. S. F., Ventura, N. R., Almeida, A. M. G. d., Garcia, L. P. F., Dias, M. d. S., and Carvalho, A. C. P. d. L. F. d. (2025). Assessing rule-based document segmentation and word normalization for legal ruling classification. In *Conference on Digital Government Research*, volume 26.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

Jegan, R. and Henrich, A. (2025). Contrasting traditional models and LLMs: An evaluation based on text segmentation. In Wartena, C. and Heid, U., editors, *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 274–281, Hannover, Germany. HsH Applied Academics.

Krassovitskiy, A., Mussabayev, R., and Yakunin, K. (2025). Llm-enhanced semantic text segmentation. *Applied Sciences*, 15(19).

Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., and Jagadish, H. V. (2008). Regular expression learning for information extraction. In Lapata, M. and Ng, H. T., editors, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 21–30, Honolulu, Hawaii. Association for Computational Linguistics.

Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., Sun, L., and Wu, H. (2022). Unified structure generation for universal information extraction. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Ollama (2024). Ollama: Run large language models locally. `https://ollama.com/`. Accessed: 25 fev. 2026.

Ollama Library (2024). Deepseek-r1 8b. `https://ollama.com/library/deepseek-r1:8b`. Large Language Model. Acesso em: 25 fev. 2026.

Queiroz, B., Cavalcante, R., and Claro, D. (2023). Desafios da tarefa de extração de informação aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 388–392, Porto Alegre, RS, Brasil. SBC.

Tang, X., Zong, Y., Phang, J., Zhao, Y., Zhou, W., Cohan, A., and Gerstein, M. (2024). Struc-bench: Are large language models good at generating complex structured tabular data? In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 12–34, Mexico City, Mexico. Association for Computational Linguistics.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. (2024). Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6):186357.