Towards a Universal Dependencies Corpus for Portuguese Epidemiological Reports

Christian Freitas[1], Livy Real[2,3], Lilian Berton[1], Valeria de Paiva[4]

[1]Universidade Federal de São Paulo, São Paulo, Brazil
[2]Universidade Federal do Amazonas, Manaus, Brazil
[3]Instituto Kunumi, Belo Horizonte, Brazil
[4]Topos Institute, Berkeley, USA

christian.freitas@unifesp.br · livy@kunumi.com · lberton@unifesp.br · valeria@topos.institute

Abstract

We present an ongoing research project focused on the construction of a Universal Dependencies (UD) corpus of Portuguese epidemiological reports derived from documents published within the Brazilian public health system. We describe findings and challenges to build such a corpus from PDF reports processed through a controlled document extraction pipeline that contrasts layout-aware extraction with raw PDF text extraction, explicitly addressing the impact of tabular content on downstream syntactic analysis. Narrative text is annotated using multiple UD parsers for Portuguese, including widely used and state-of-the-art tools, and their outputs are systematically compared using descriptive structural indicators and targeted qualitative inspection.

Our analysis highlights domain-specific challenges in epidemiological texts and shows that document extraction and representation choices have a stronger effect on parsing behavior than parser selection alone. Based on these findings, we identify robust preprocessing configurations and discuss design choices for a UD-epidemiological corpus to support future research on syntactic parsing, domain adaptation, and downstream natural language processing tasks in epidemiology and public health.

1 Introduction

Universal Dependencies (UD) has become a widely adopted framework for syntactic annotation, enabling cross-linguistic consistency and facilitating the development and evaluation of dependency parsers across languages and domains (Nivre et al., 2016, 2020). For Portuguese, several UD treebanks and parsing tools have been developed, supporting a wide range of natural language processing applications (Rademaker et al., 2017; Branco et al., 2022; Sanches Duran et al., 2025). Despite these advances, most existing resources and evaluations focus on newswire or general-domain texts, leaving specialized domains comparatively underexplored. Exceptions include Di Felippo et al. (2024); Souza and Freitas (2023).

Epidemiological reports constitute a particularly challenging domain for syntactic parsing. Such documents typically combine technical terminology, numerical expressions, abbreviated forms, and complex syntactic constructions, often embedded in heterogeneous document layouts that include tables, lists, and scanned pages. These characteristics can negatively impact both text extraction and downstream syntactic analysis, especially when models trained on general-domain data are applied without adaptation.

We intend to address this gap by producing a Universal Dependencies corpus of Portuguese epidemiological reports derived from documents published within the Brazilian public health surveillance system. In this work, we focus on a preliminary step for the construction of the corpus through a controlled document processing pipeline that compares different text extraction strategies, evaluates the impact of optical character recognition when required, and applies multiple UD parsers for Portuguese, including state-of-the-art and widely used tools. Our goal is not to propose new parsing models, but rather to quantify parsing behavior and common error patterns in this specialized domain and to release a curated corpus that can support future research on syntactic analysis and domain adaptation. Our long-term goal is to produce a reliable UD corpus for epidemiological reports in Portuguese.

2 Epidemiological Reports

The SIREVA (Sistema Regional de Vacinas) system is a public health surveillance initiative coordinated in Brazil within the Unified Health System in Portuguese, Sistema Único de Saúde (SUS). The SIREVA-SUS system focuses on the monitoring of invasive bacterial diseases and vaccine-preventable

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 228–237

pathogens. The system produces periodic epidemiological reports that consolidate laboratory-confirmed cases, serotype distributions, and temporal and geographic trends, serving as an important source of information for epidemiological understanding and public health decision-making.

The documents analyzed in this work consist of official SIREVA-SUS epidemiological reports, which are published in PDF format. These reports typically combine continuous narrative text with tables, lists, and summary statistics. They may include scanned pages depending on the publication year and source. From a natural language processing perspective, this heterogeneous structure poses challenges for automatic text extraction and syntactic analysis, as layout artifacts and domain-specific formatting can negatively affect tokenization, sentence segmentation, and dependency parsing. Figure 1 illustrates this contrast using content extracted directly from the 2024 SIREVA-SUS report: a representative structured data table alongside its corresponding narrative interpretation.

Tabela 1. Número de isolados invasivos por grupo etário e sexo

Grupo etário	*Sexo*							
	Masculino		Feminino		Sem dado		Total	
	n	%	n	%	n	%	n	%
< 12 meses	25	52,1	21	43,8	2	4,2	48	18,1
12–23 meses	11	57,9	8	42,1	0	0,0	19	7,2
24–59 meses	17	56,7	13	43,3	0	0,0	30	11,3
Subtotal (1)	**53**	**54,6**	**42**	**43,3**	**2**	**2,1**	**97**	**36,6**
5–14 anos	13	46,4	15	53,6	0	0,0	28	10,6
15–29 anos	14	66,7	7	33,3	0	0,0	21	7,9
30–49 anos	18	56,3	14	43,8	0	0,0	32	12,1
Subtotal (2)	**45**	**55,6**	**36**	**44,4**	**0**	**0,0**	**81**	**30,6**
50–59 anos	10	40,0	15	60,0	0	0,0	25	9,4
≥ 60 anos	26	41,9	36	58,1	0	0,0	62	23,4
Subtotal (3)	**36**	**41,4**	**51**	**58,6**	**0**	**0,0**	**87**	**32,8**
Total	**134**	**50,6**	**129**	**48,7**	**2**	**0,8**	**265**	**100,0**

*"Do total de **265** amostras H. influenzae: **166** se referem à cultura e **99** se referem a PCR em tempo real."*

Figure 1: Example of content from the 2024 SIREVA-SUS report: a structured data table (top) followed by its narrative interpretation (bottom). This juxtaposition illustrates the heterogeneous nature of epidemiological documents, where tabular evidence and textual claims coexist and must be handled separately by the processing pipeline.

In this study, the SIREVA-SUS reports serve as a representative example of real-world epidemiological documents in Portuguese. By focusing on this data source, we aim to evaluate the behavior of UD parsers under domain-specific conditions and to construct a corpus that reflects the linguistic and structural characteristics commonly found in epidemiological surveillance reports.

Table 1: Pages and heuristically detected tables per SIREVA-SUS report.

Year	**Pages**	**Tables**
2013	42	51
2014	41	53
2015	43	51
2016	43	89
2017	41	62
2018	41	61
2019	38	61
2020	36	52
2021	37	53
2022	37	58
2023	43	69
2024	42	51
Total	**484**	**711**

The documents analyzed in this work consist of official SIREVA-SUS epidemiological reports, which are published in PDF format and made publicly available by the Adolfo Lutz Institute.[1]

3 Methodology

Before constructing the full corpus, we adopt a pilot-based evaluation strategy in which the entire document processing and parsing pipeline is applied to a single, representative epidemiological report. Specifically, all experiments reported in this section are conducted using the SIREVA-SUS report from 2024.

This document was selected because it is structurally representative of the collection as a whole, combining narrative text, extensive tabular content, and modern PDF formatting. By focusing initially on a single report, we are able to analyze the effects of document extraction choices, table handling, and parser behavior in a controlled setting, while avoiding confounding variation introduced by inter-document heterogeneity.

We therefore adopt a representation strategy that separates text and tables *physically*, while preserving their *logical connections* at the document level. Conceptually, each report is modeled as a collection of textual statements and tabular evidence objects, linked by explicit relations that capture their rhetorical and evidential roles.

[1] https://www.ial.sp.gov.br/ial/publicacoes/boletim

3.1 Text and Table Extraction

During preprocessing, narrative text blocks (e.g., paragraphs, section summaries) and tables are extracted independently from the original PDF documents. Text blocks are segmented into coherent units (typically paragraphs) and assigned stable identifiers. Tables are parsed into structured representations that preserve row and column headers, cell values, units, captions, and footnotes. Numeric values are retained in their original form and are not subjected to language modeling.

This separation ensures that tables remain amenable to deterministic processing, validation, and normalization, while text blocks remain suitable for natural language processing techniques.

3.2 Tables as Structured Evidence

Each table is treated as a structured evidence object rather than as a textual artifact. Rows and columns are interpreted according to their semantic roles (e.g., disease, year, geographic region, measure type), and individual cell entries correspond to atomic factual statements. Captions and table-local annotations are preserved as metadata, as they often specify measurement conventions, exclusions, or temporal scope.

By maintaining tables in a structured form, the pipeline supports downstream tasks such as unit normalization, consistency checking, aggregation, and cross-document comparison, which are essential in medical and epidemiological settings.

3.3 Linking Textual Claims and Tables

To capture the intended relationship between narrative and data, explicit links are introduced between text blocks and tables.These links are inferred automatically using a combination of explicit textual references (e.g., "Table 3 shows..."), document structure, proximity heuristics, and caption semantics. Each link is labeled with one of three coarse-grained relation types: *supported_by* (the narrative claim is directly backed by tabular data), *elaborates* (the text expands on information presented in a table), and *summarizes* (the text condenses tabular content into a higher-level statement). Manual validation of these automatically inferred links is planned as part of the corpus curation process described in Section 7.

This yields a document-level representation in which textual claims and tabular evidence are connected but remain distinct. Such a representation makes it possible to trace which quantitative data support which assertions, to identify uncited or weakly supported claims, and to reason jointly over text and data without conflating their roles.

3.4 Implications for Information Extraction

Under this representation, information extraction from tables is performed deterministically over structured data, while UD processing is applied primarily to narrative text, captions, and annotations. Large language models, when used, are restricted to interpreting metadata and mapping textual descriptions to canonical schemas, and are never treated as the source of numeric ground truth.

This separation-of-concerns design reduces error propagation, supports validation and provenance tracking, and reflects the epistemic distinction between quantitative evidence and its narrative interpretation. In oversight and reporting contexts, this distinction is critical for ensuring transparency and analytical reliability.

4 Parsing Configurations and Tool Selection

Why Universal Dependencies. We adopt the Universal Dependencies (UD) framework as the syntactic representation for this corpus for three reasons: cross-resource comparability, parser diversity, and structural transparency. First, UD provides a linguistically motivated but application-agnostic annotation scheme that is consistent across languages and domains, allowing the resulting corpus to be compared directly with existing Portuguese treebanks and with epidemiological corpora in other languages. Second, most widely used dependency parsers for Portuguese either natively produce UD annotations or can be reliably mapped to UD, making it possible to evaluate multiple parsing configurations within a shared representational space. Finally, UD's explicit encoding of predicate–argument structure, modification, coordination, and clause boundaries makes it well suited for analyzing the syntactic realizations of epidemiological statements, such as causal claims, temporal descriptions, and quantified assertions. Our goal is not to advance syntactic theory, but to obtain a stable, interpretable syntactic layer that supports error analysis and future downstream tasks, including information extraction and argument-level modeling.

Why CoNLL-U. All parsed outputs are normalized to the CoNLL-U format, the official in-

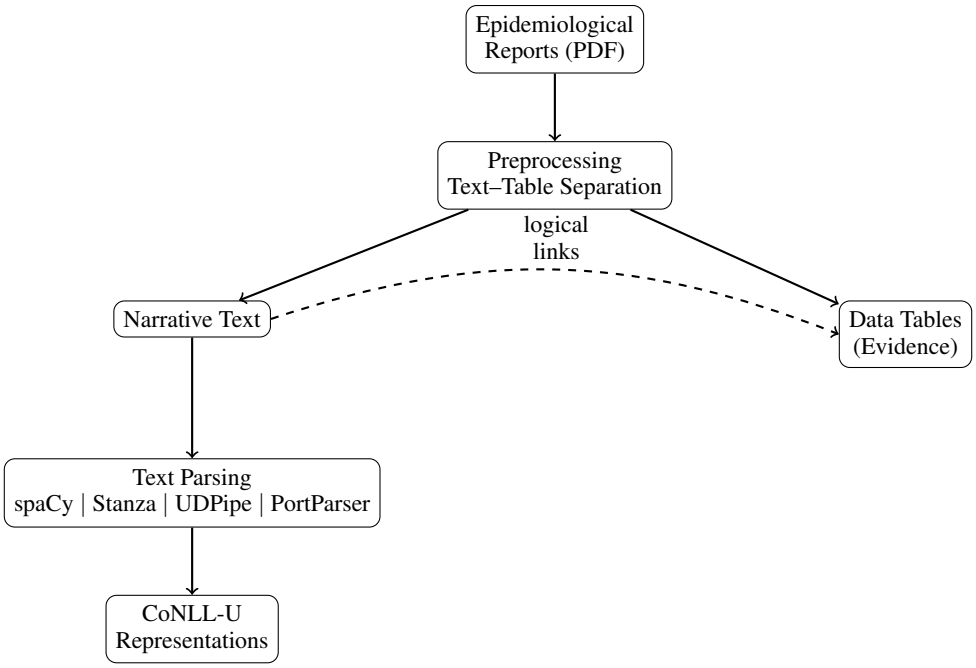

Figure 2: Processing pipeline for epidemiological reports. Reports are preprocessed to separate narrative text from tabular data while preserving their logical connections. Narrative text is parsed using multiple dependency parsers and converted to CoNLL-U format, while tables are retained as structured evidence objects and excluded from linguistic parsing.

terchange format of the Universal Dependencies project. CoNLL-U provides a compact but expressive representation that integrates tokenization, lemmatization, morphological features, part-of-speech tags, and dependency relations in a single, line-oriented structure. Using CoNLL-U enables direct comparison between parsers with different internal architectures and preprocessing strategies, while preserving sufficient linguistic detail for qualitative inspection and downstream reuse. In addition, CoNLL-U supports partial annotation, sentence-level alignment, and post-hoc correction, which is essential when working with automatically extracted text from noisy PDF sources. By committing to CoNLL-U, we ensure that the released corpus can be easily inspected, extended, or re-parsed as improved models or domain-adapted parsers become available.

5 Document Extraction with Docling

Docling (IBM Research, 2024) is a document processing framework designed to extract structured textual content from complex PDF documents, with particular emphasis on preserving layout information and separating different content types, such as continuous text and tables. Unlike raw PDF text extraction methods, which operate primarily at the character or line level, Docling performs layout-aware segmentation, allowing for a more faithful reconstruction of document structure.

In the context of epidemiological reports, which frequently combine narrative descriptions with tabular data and heterogeneous formatting, layout-aware extraction is especially relevant. By explicitly distinguishing between running text and tables, Docling enables downstream linguistic processing to focus on syntactically meaningful textual units, while avoiding artifacts commonly introduced by table structures when treated as plain text.

In this work, Docling is used as the primary document extraction method and is compared against a raw PDF text extraction baseline. This comparison allows us to assess the impact of layout-aware document processing on subsequent UD parsing quality and to quantify the extent to which document structure influences syntactic analysis in epidemiological texts.

5.1 Removal of Markdown Structural Punctuation in Docling Output

During initial experiments, we observed that Docling exports textual content from PDF documents using a Markdown-based representation, particularly when encoding tabular structures. In this representation, structural elements such as vertical

bars (|), header markers (#), and column separators become part of the linearized text.

This behavior had a direct impact on downstream analyses. Markdown-specific characters were tokenized by the linguistic parsers (spaCy, Stanza, and UDPipe), leading to an artificial inflation in the number of tokens, lexical types, and sentences. As a consequence, medically relevant terms (e.g., *meningite*, *pneumonia*) were repeatedly counted in contexts that do not correspond to natural language usage, but rather to the structural encoding of tables.

To mitigate this effect and ensure comparability with other PDF text extraction methods (such as PyPDF and PyMuPDF), we introduced an explicit post-processing step for text extracted with Docling. The goal of this step was to preserve the informational content conveyed by tables while eliminating Markdown-specific structural punctuation that does not carry linguistic meaning.

Adopted Strategy. The adopted strategy consists of removing typical Markdown characters used for table and title formatting, including vertical bars, header markers, column separator lines (e.g., `---`), and redundant punctuation introduced by the Markdown layout. Only the textual and numerical content of table cells is retained, in a linearized form comparable to the output produced by traditional PDF text extractors.

Comparative analyses across extraction variants show that the original Docling output with Markdown markup yields substantially higher token and sentence counts. After the removal of Markdown structural punctuation, these values closely align with those observed for other PDF readers. This confirms that the observed explosion in token and sentence counts is not driven by semantic content, but by the structural representation of tables.

These findings indicate that current linguistic parsers are not designed to operate directly on Markdown-encoded tabular structures, reinforcing the need for careful preprocessing when technical documents rich in tables are used as input to syntactic parsers.

6 UD Parsers

Given the choice of Universal Dependencies as the syntactic framework, the selection of parsing tools follows naturally. We focus on parsers that (i) natively produce UD-compliant analyses, (ii) support Portuguese with publicly available models, and (iii) differ substantially in architectural design, training data, and intended use. This allows us to examine how distinct parsing paradigms behave when applied to epidemiological text extracted from complex PDF documents, while holding the annotation scheme and output format fixed. By normalizing all parser outputs to CoNLL-U, we ensure that observed differences reflect genuine parsing behavior rather than representational incompatibilities. The resulting comparison is therefore not a competition between systems, but a controlled evaluation of robustness, error patterns, and domain sensitivity under a shared syntactic standard.

To assess the robustness of dependency parsing in the epidemiological domain, we employ three widely used UD parsers for Portuguese: spaCy, Stanza, and UDPipe. These tools represent different design choices and levels of linguistic modeling, and are commonly adopted as baselines in dependency parsing evaluations. In addition, we include the PortParser (Lopes and Pardo, 2024), a state-of-the-art dependency parser for Portuguese.

spaCy. spaCy is an industrial-strength natural language processing library that provides efficient pipelines for tokenization, part-of-speech tagging, and dependency parsing (Honnibal et al., 2020). Its dependency parser is based on transition-based neural models optimized for speed and scalability. Although spaCy is not primarily designed for linguistic research, it supports Universal Dependencies labels and is frequently used in applied NLP settings. In this work, spaCy serves as a pragmatic baseline, allowing us to evaluate how a general-purpose, high-performance parser behaves when applied to specialized epidemiological texts. spacy offers a library to produce Universal Dependencies in a CNLL format. Note though that spaCy was developed for English and its Portuguese models are not particularly fine-tuned for medical data.

Stanza. Stanza (Qi et al., 2020) is a neural NLP toolkit developed by the Stanford NLP Group, designed with a strong focus on linguistic accuracy and multilingual support. It provides end-to-end pipelines for tokenization, morphological analysis, part-of-speech tagging, lemmatization, and dependency parsing, all trained within the Universal Dependencies framework. Stanza's models rely on deep contextualized representations and have shown competitive performance across multiple languages and treebanks.

UDPipe. UDPipe (Straka et al., 2016) is a trainable pipeline for processing text in the CoNLL-U format, offering models for tokenization, tagging, lemmatization, and dependency parsing. It has been extensively used in shared tasks and benchmark studies related to UD, and is known for its efficiency and reproducibility. UDPipe models are trained directly on UD treebanks and follow the official annotation guidelines closely, making the tool particularly suitable for comparative evaluations. In this study, UDPipe provides a strong and well-established baseline for assessing parsing quality in Portuguese epidemiological reports.

PortParser. PortParser (Lopes and Pardo, 2024) is specifically designed for Portuguese and leverages recent advances in neural dependency parsing, achieving top performance on standard Portuguese UD benchmarks. Its architecture and training strategy are optimized to capture language-specific syntactic phenomena that are often underrepresented in multilingual or general-purpose models.

The inclusion of PortParser allows us to establish a reference for parsing quality in Portuguese and to assess how models behave when applied to a specialized and out-of-domain setting such as epidemiological reports. By comparing PortParser with more general-purpose UD parsers, we aim to identify whether gains observed in benchmark evaluations transfer to real-world epidemiological texts, which exhibit domain-specific terminology, numerical expressions, and heterogeneous document structures.

In this section, we analyze the behavior of all parsers (spaCy, Stanza, UDPipe, and PortParser v2) across different text extraction scenarios using basic structural metrics: number of sentences, number of tokens, number of word-form types, and number of lemma types. Rather than ranking parsers by performance, our objective is to ground the discussion in observed quantitative differences and to understand how parser design choices interact with document preprocessing decisions in the syntactic analysis of epidemiological reports.

Table 2 summarizes the main structural statistics for each combination of extraction scenario and parser, providing the empirical basis for the analyses discussed below.

Text extraction scenarios. Table 2 summarizes four text extraction settings designed to isolate the effects of (i) raw PDF extraction versus layout-aware extraction and (ii) the presence of linearized tabular content in the parser input. We use the following scenario labels throughout the results: **A_raw_pypdf** (raw text extracted with PyPDF), **D_raw_pymupdf** (raw text extracted with PyMuPDF), **B_docling_text_only** (Docling layout-aware extraction that still retains linearized table content), and **B2_docling_text_only_no_tables** (Docling extraction with explicit removal of tables, keeping only running narrative text).

Lexical stability (tokens and types). As shown in Table 2, spaCy, Stanza, and PortParser v2 exhibit highly similar behavior with respect to the total number of tokens, word-form types, and lemma types under raw extraction scenarios (PyPDF and PyMuPDF). In these settings, all three parsers converge to nearly identical values, indicating stable tokenization and lemmatization when the input text does not contain complex tabular structures or artificial markup.

UDPipe, in contrast, displays a markedly different lemmatization profile. Although its token counts and word-form type counts are comparable to those of the other parsers, the number of distinct lemmas is consistently lower across all scenarios. This pattern, visible in both raw and Docling-based extractions, suggests a more aggressive normalization strategy or a less fine-grained lemmatization process, which may impact downstream tasks that depend on lexical diversity.

Sentence segmentation variability. Sentence segmentation shows the largest variation across parsers. For identical input text, UDPipe consistently produces the highest number of sentences, while spaCy yields intermediate values and Stanza produces fewer sentences. PortParser v2 exhibits the most conservative segmentation behavior, generating the smallest number of sentences in most scenarios (Table 2).

These differences directly affect the granularity of linguistic units and have implications for downstream tasks such as information extraction, text–table alignment, discourse analysis, and the modeling of long-range syntactic and semantic relations.

Impact of text extraction and preprocessing. Beyond parser-specific behavior, Table 2 shows that the choice of text extraction method exerts an effect that is comparable to, and in some cases greater than, the choice of parser. In the Docling extraction scenario that retains linearized ta-

Table 2: Descriptive statistics of parsed outputs across document extraction variants and dependency parsers, including PortParser v2.

Scenario	Parser	Sentences	Tokens	Types (form)	Types (lemma)
A_raw_pypdf	PortParser v2	276	12193	1773	1747
A_raw_pypdf	spaCy	724	11347	1762	1732
A_raw_pypdf	Stanza	471	11297	1775	1755
A_raw_pypdf	UDPipe	1273	12405	1781	833
B2_docling_text_only_no_tables	PortParser v2	95	1603	363	346
B2_docling_text_only_no_tables	spaCy	145	1559	370	348
B2_docling_text_only_no_tables	Stanza	172	1575	356	341
B2_docling_text_only_no_tables	UDPipe	163	1607	354	188
B_docling_text_only	PortParser v2	1115	28067	1787	1758
B_docling_text_only	spaCy	2304	21477	1805	1774
B_docling_text_only	Stanza	290	21420	1830	1810
B_docling_text_only	UDPipe	1689	23114	1841	817
D_raw_pymupdf	PortParser v2	276	12186	1771	1745
D_raw_pymupdf	spaCy	648	11334	1759	1729
D_raw_pymupdf	Stanza	480	11284	1773	1753
D_raw_pymupdf	UDPipe	1294	12395	1779	831

bles, all parsers exhibit a substantial increase in token counts and a pronounced inflation in sentence counts. This effect is particularly evident for spaCy and UDPipe, whose sentence counts increase by an order of magnitude relative to cleaner extraction settings.

When tables are explicitly removed, (`docling_text_only_no_tables`), parser behavior returns to patterns closely aligned with those observed under raw PDF extraction. This confirms that the degradation in parsing stability is driven not by the narrative content itself, but by the representational form of tabular data.

This effect arises from the presence of linearized tabular content, which introduces non-natural patterns of punctuation, repetition, and structural markers that interfere with both tokenization and sentence boundary detection. When tables are explicitly removed, parser behavior returns to patterns similar to those observed under raw PDF extraction. This confirms that it is not the narrative text itself that degrades parsing behavior, but rather the representational form of tabular data.

Sentence fragmentation under noisy extraction. Figure 3 complements Table 2 by normalizing sentence counts per 1k tokens. This visualization highlights that extraction pipelines including linearized tables consistently lead to higher sentence fragmentation across all parsers. Although the magnitude of the effect varies by parser, the overall trend is stable: noisy extraction amplifies segmentation artifacts independently of parser architecture.

Morphosyntactic profile and structural noise. Beyond sentence-level effects, the distribution of morphosyntactic categories further reflects the impact of document representation. As illustrated in Figure 4, scenarios with linearized tables exhibit elevated proportions of punctuation and symbol tokens, corresponding to layout markers and separators rather than linguistic structure.

When tables are removed, morphosyntactic profiles stabilize across parsers. The relative proportions of core categories such as nouns, verbs, and adjectives become consistent, indicating that narrative epidemiological text presents a regular grammatical structure once freed from tabular noise. This pattern holds across both general-purpose parsers and PortParser v2, reinforcing the conclusion that preprocessing choices dominate morphosyntactic behavior.

Summary. Taken together, the results demonstrate that no parser evaluated in this study is robust to the naive linearization of tables, that sentence segmentation varies substantially across parsers even for identical input text, and that differences in lemmatization, particularly in UDPipe, may affect tasks sensitive to lexical diversity. Crucially, the

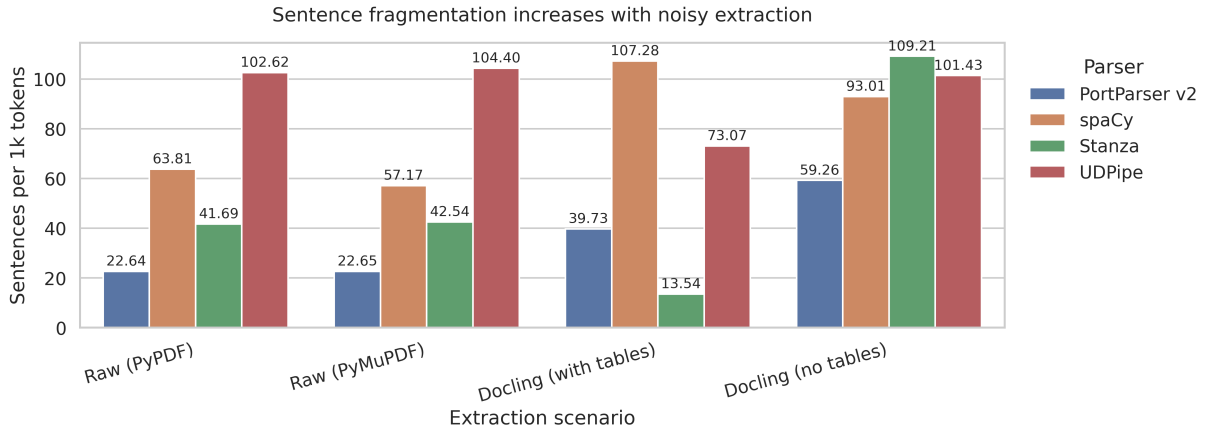

Figure 3: Sentence fragmentation across extraction scenarios, measured as sentences per 1k tokens. **Lower values indicate more stable sentence boundary detection.** Noisy extraction pipelines with linearized tables substantially increase fragmentation across all parsers, while explicit table removal (B2) yields values comparable to raw PDF extraction, suggesting more linguistically plausible segmentation.

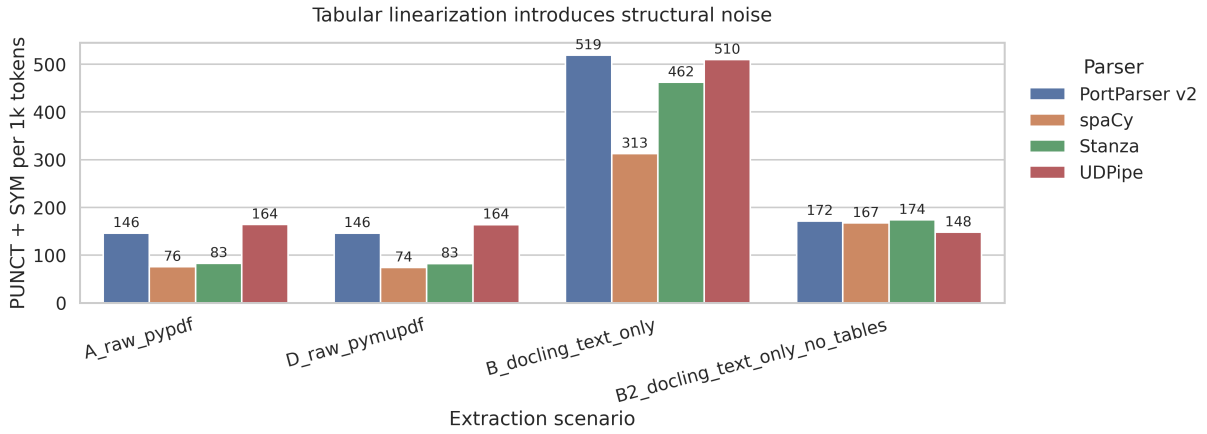

Figure 4: Structural noise across extraction scenarios and parsers, measured as the number of punctuation and symbol tokens per 1k tokens. **Lower values reflect a cleaner morphosyntactic profile with fewer non-linguistic artifacts.** Linearized tables substantially increase structural noise across all parsers, while explicit table removal (B2) yields profiles consistent with raw PDF extraction.

quantitative evidence shows that document preprocessing and representation choices are at least as critical as parser selection, underscoring the need for pipelines that explicitly preserve the distinction between narrative text and tabular data in epidemiological reports.

7 Conclusion and Future Work

This paper presented an ongoing effort to construct a Universal Dependencies corpus for Portuguese epidemiological reports derived from documents published within the Brazilian public health system. As an initial and deliberately controlled step, the analysis focused on a single, representative SIREVA-SUS report from 2024, which was used as a pilot document to evaluate the complete document extraction and parsing pipeline. Code and resources from this paper are available at the project repository[2].

The results show that document extraction and representation choices exert a strong influence on syntactic parsing behavior, even when analysis is restricted to a single report. In particular, raw PDF extraction and the linearization of tabular content lead to substantial inflation in sentence counts and structural artifacts across all parsers considered. In contrast, layout-aware extraction combined with explicit table removal produces more stable and linguistically plausible inputs, reducing segmentation noise and yielding more consistent parsing

[2] https://github.com/ChristianSF/SIREVA-SUS-Corpus

statistics. These effects are observed across both general-purpose parsers and a parser specifically designed for Portuguese, indicating that preprocessing decisions outweigh parser-specific differences in this domain, where semi-structured content is central and the original data is available exclusively in PDF format.

The comparative evaluation further highlights systematic differences in sentence segmentation and lemmatization strategies across parsers, with direct implications for downstream tasks such as information extraction, text–table alignment, and discourse-level analysis. Taken together, the findings reinforce the need for document processing pipelines that explicitly preserve the distinction between narrative text and structured data when building syntactically annotated resources from real-world technical documents.

As future work, we will extend the validated pipeline to the full collection of SIREVA-SUS epidemiological reports, moving beyond the single-document pilot to support broader generalization of the findings reported here. The complete corpus will be released together with detailed documentation and reproducible preprocessing scripts, enabling reuse for research on syntactic parsing, domain adaptation, and downstream NLP tasks in epidemiology and public health.

A key next step is the introduction of systematic quantitative evaluation. We plan to manually annotate a small but representative subset of the corpus to serve as a gold standard, enabling the computation of standard dependency parsing metrics such as Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS). This will allow direct comparison of parser behavior in the epidemiological domain against benchmark results on general-domain Portuguese treebanks, complementing the structural indicators used in the current pilot.

Finally, future work will detail the manual annotation and curation plan for the full UD corpus. This includes defining domain-specific annotation guidelines to handle constructions that are frequent in epidemiological reports but underrepresented in existing Portuguese UD treebanks, such as abbreviated nominal phrases, table captions used as standalone sentences, and numerical expressions with embedded units and uncertainty markers.

8 Declaration of Generative AI in the writing process

During the preparation of this work, the authors used ChatGPT and Gemini in order to improve the language, grammar, and flow of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

António Branco, João Ricardo Silva, Luís Gomes, and João António Rodrigues. 2022. Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5617–5626, Marseille, France. European Language Resources Association.

Ariani Di Felippo, Norton Trevisan Roman, Thiago Alexandre Salgueiro Pardo, and Lucas Panta de Moura. 2024. The dantestocks corpus: an analysis of the distribution of universal dependencies-based part-of-speech tags. *Revista da ABRALIN*, 22(2):249–271.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

IBM Research. 2024. Docling: Layout-aware document processing for complex pdfs. `https://github.com/DS4SD/docling`. Accessed: 2025-01.

Lucelene Lopes and Thiago Pardo. 2024. Towards portparser-a highly accurate parsing system for brazilian portuguese following the universal dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 401–410.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online. Association for Computational Linguistics.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria De Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the fourth international conference on dependency linguistics (Depling 2017)*, pages 197–206.

Magali Sanches Duran, Elvis A. de Souza, Maria das Graças Volpe Nunes, Adriana Silvina Pagano, and Thiago A. S. Pardo. 2025. Extending the enhanced Universal Dependencies – addressing subjects in pro-drop languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 143–152, Ljubljana, Slovenia. Association for Computational Linguistics.

Elvis Souza and Claudia Freitas. 2023. Annotation of fixed multiword expressions (MWEs) in a Portuguese Universal Dependencies (UD) treebank: Gathering candidates from three different sources. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 442–450, Belo Horizonte, Brazil. Association for Computational Linguistics.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).