

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Text Segmentation of City Council Minutes in European Portuguese

José Miguel Moreira Isidro



Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Ricardo Campos

Second Supervisor: Prof. Sérgio Nunes

July 13, 2026

Text Segmentation of City Council Minutes in European Portuguese

José Miguel Moreira Isidro

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Carla Teixeira Lopes

Examiner: Prof. David Semedo

Member: Prof. Ricardo Campos

July 13, 2026

Resumo

Com a crescente aposta na transparência pública, as câmaras municipais têm vindo a converter os seus arquivos e atas em bases de dados digitais. No entanto, a consulta manual destes documentos continua a ser difícil, pois são textos longos e densos, organizados segundo convenções administrativas que variam entre câmaras. Esta dissertação foca-se na segmentação de texto de atas de câmaras municipais, com o objetivo de dividir automaticamente estes registos em temas ou assuntos coerentes. Estas atas são um desafio para a computação porque são muito extensas, utilizam uma linguagem administrativa específica e nem sempre recorrem a marcadores ou separadores normalizados que indiquem claramente a transição entre assuntos, o que as distingue de outros tipos de texto, como notícias ou artigos científicos.

Para resolver este problema, propomos o SegNSP, uma abordagem que dispensa dados anotados com etiquetas semânticas, adaptando o objetivo de pré-treino Next Sentence Prediction para detetar as fronteiras entre tópicos, através da continuidade entre frases adjacentes. A metodologia introduz melhorias significativas face às arquiteturas baseadas em *Transformers*, designadamente através de rácios de amostragem equilibrados, estratégias de amostragem negativa robusta e uma função de perda dedicada à segmentação.

A avaliação do modelo recorre a várias métricas de precisão, nomeadamente P_k , WindowDiff (WD), Boundary F_β -measure ($B-F_1$) e Boundary Similarity (B), aplicadas a dois conjuntos de dados: o conjunto de referência internacional WikiSection e uma versão adaptada da nova coleção CitiLink-Minutes. Destaca-se a nossa contribuição no desenvolvimento deste último, que constitui a primeira coleção de referência para atas administrativas em português europeu, incluindo 120 documentos manualmente anotados de seis municípios distintos.

Os resultados demonstram a eficácia do SegNSP, que estabelece uma base de referência linear competitiva, alcançando um $B-F_1$ de 79.22 e um B de 59.31 no CitiLink-Minutes, superando vários modelos neuronais e não supervisionados. Para colmatar as limitações da segmentação linear em documentos com estruturas aninhadas, introduzimos de seguida o HSeg, um modelo hierárquico, recursivo e generativo baseado em modelos de linguagem de grande escala (Large Language Models (**LLMs**)). Avaliado tanto no domínio administrativo como no corpus WIKI-727K, o HSeg supera o estado da arte reportado na literatura. No CitiLink-Minutes, o modelo atinge um $B-F_1$ de 89.16 na extração de assuntos recorrendo a modelos proprietários de última geração, abrindo também caminho para a validação de alternativas *open-weight* especializadas. Adicionalmente, no corpus WIKI-727K, demonstra que mesmo **LLMs** otimizados e mais leves conseguem superar as arquiteturas dedicadas anteriores com uma melhoria absoluta de +9,06 pontos na métrica F_1 (80,66 face a 71,60), com modelos de última geração como o Claude Opus 4.7 a estabelecerem ainda os melhores resultados em termos de P_k e precisão. Em última análise, este trabalho procura tornar as atas municipais mais acessíveis e úteis, tanto para funcionários públicos como para cidadãos.

Palavras-chave: Processamento de Linguagem Natural • Segmentação de Texto • LLMs • CitiLink • SegNSP • HSeg • Atas de Câmaras Municipais • Português Europeu •

Abstract

Municipal archives have evolved into vast digital repositories, driven by legal requirements for transparency. However, manual information discovery within these documents remains difficult, as they are lengthy, dense texts organized according to administrative conventions that vary across municipalities. This dissertation addresses the challenge of navigating these records by exploring text segmentation, the task of dividing documents into topically coherent sections, specifically within the domain of administrative minutes. City council minutes present unique challenges, including the use of specialized procedural jargon, the inconsistent use of standardized markers or separators that clearly indicate thematic transitions, and the presence of interleaved discussions, distinguishing them from standard news or scientific corpora.

We propose SegNSP, a label-agnostic approach that adapts the Next Sentence Prediction pre-training objective to identify topic boundaries. Our methodology enhances standard transformer-based architectures with domain-balanced ratios, harder negative sampling, and a segmentation-aware loss function to capture subtle thematic transitions in noisy administrative discourse.

Evaluation is conducted using a multi-dimensional framework, including P_k , WindowDiff (WD), Boundary F_β -measure ($B-F_1$), and Boundary Similarity (B), across two primary datasets: the well-established WikiSection benchmark and an adapted version of the newly created CitiLink-Minutes. We contributed to the development of the latter, which constitutes the first European Portuguese dataset for administrative text segmentation, comprising 120 manually annotated documents from six municipalities.

Results indicate that SegNSP establishes a competitive linear baseline, reaching a $B-F_1$ of 79.22 and a B of 59.31 on the CitiLink-Minutes corpus, outperforming several neural and unsupervised models. To address the limitations of linear segmentation on deeply nested documents, we further introduce HSeg, a recursive, generative hierarchical framework powered by Large Language Models (**LLMs**). Evaluated on both the administrative domain and the large-scale, open-domain WIKI-727K benchmark, HSeg surpasses the state of the art reported in the literature. On the CitiLink-Minutes corpus, it achieves a micro-level $B-F_1$ of 89.16 utilizing frontier proprietary models, while concurrently facilitating the evaluation of domain-specialized open-weight alternatives. Furthermore, on the WIKI-727K benchmark, it decisively outperforms prior dedicated CRF architectures with an absolute improvement of +9.06 F_1 points (80.66 vs. 71.60), demonstrating the exceptional retrieval capabilities of even lightweight, cost-effective models, while frontier models such as Claude Opus 4.7 separately establish the best results in terms of P_k and precision. Ultimately, this project aims to make municipal minutes more accessible and useful, both for public officials and for citizens.

Keywords: Natural Language Processing • Text Segmentation • LLMs • CitiLink • SegNSP • HSeg • City Council Minutes • European Portuguese •

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for achieving a better and more sustainable future for all. It includes 17 goals (<https://sdgs.un.org/goals>) to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

This dissertation, which focuses on the automated Text Segmentation (TS) and structural parsing of city council minutes, directly contributes to these global objectives. By developing open-source Natural Language Processing (NLP) architectures that transform dense, semi-structured administrative records into accessible, highly navigable civic data, this research supports the advancement of institutional transparency and citizen engagement in local governance.

The specific SDGs most closely related to this work are:

SDG 11 Make cities and human settlements inclusive, safe, resilient and sustainable

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

SDG	Target	Contribution	Performance Indicators and Metrics
11	11.3	By making complex city council deliberations transparent and easily searchable via the CitiLink platform, this work facilitates greater citizen participation and integrated human settlement planning.	Number of municipalities integrated into civic tech platforms; Active user engagement metrics on the portal.
16	16.6	The text segmentation models (SegNSP and HSeg) automate the structuring of public records, helping municipal bodies operate as more effective, accountable, and transparent institutions.	Segmentation accuracy (e.g., B-F ₁ scores); Reduction in time required to retrieve specific municipal records.
	16.10	The release of open-source research artifacts, such as the CitiLink-Minutes dataset, provides the foundational data necessary to ensure public access to government information.	Number of open-source dataset downloads; Integration rate into downstream retrieval-augmented generation (RAG) systems.

Acknowledgements

I would like to express my deepest gratitude to my family, especially my mother and my sister, for their unconditional love, encouragement, and unwavering support throughout the years. Their belief in me has been a constant source of strength, and I would not be where I am today without them.

To my dearest friends, thank you for always being there when I needed you, whether to offer guidance, support, or occasionally lead me astray on adventures that became some of my most cherished memories. The laughter, companionship, and countless moments we shared made this journey far more meaningful and enjoyable.

Finally, I would like to sincerely thank my supervisors for believing in me, challenging me, and guiding me throughout this work. Their high standards, constant encouragement, and relentless pursuit of excellence pushed me to grow not only as a researcher and professional, but also as a person. For that, I am truly grateful.

This work was funded within the scope of the project CitiLink¹, with reference 2024.07509.IACDC, which is co-funded by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, measure RE-C05-i08.M04 - "To support the launch of a programme of R&D projects geared towards the development and implementation of advanced cybersecurity, artificial intelligence and data science systems in public administration, as well as a scientific training programme," as part of the funding contract signed between the Recovering Portugal Mission Structure (EMRP) and the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology), as intermediary beneficiary.²

This work was supported by the Portuguese National Advanced Computing Network (RNCA) and the Deucalion supercomputer, which is managed by FCCN (FCCN/FCT). The computational resources were provided through the Artificial Intelligence Virtual Laboratory (AIVLAB) under the project "CitiLink Text Segmentation", with reference 2026.00028.AIVLAB.DEUCALION.

Special thanks to the AMALIA-media project at FCUP.

José Miguel Moreira Isidro

¹<https://projeto.citilink.inesctec.pt/>

²<https://doi.org/10.54499/2024.07509.IACDC>

Declaration of honour

I declare that this dissertation is my original work and has not previously been submitted or assessed in any other course or curricular unit, whether at this or any other institution.

References to other authors (statements, ideas, thoughts), as well as the use of published works fully complies with the rules of attribution and are appropriately cited in the text and bibliography, in accordance with reference standards. I am aware that the practice of plagiarism or self-plagiarism constitutes an academic offence, as established in the U.Porto Code of Ethics. I further declare that I have included text(s) published in co-authorship, and that I have specified, with transparency and rigor, the contribution of each co-author, as well as stated whether there exists any other dissertation in which the same text(s) may also have been included.

I further declare that Artificial Intelligence tools were used to assist in the writing process of this dissertation, specifically to enhance clarity and readability. The use of such tools and their results were carefully reviewed for the detection of errors, inaccuracies, unfounded attributions, or other forms of content and argument bias, as well as for the determination of authorship.

José Miguel Moreira Isidro

José Miguel Moreira Isidro

“The brightest days owe their meaning to the darker ones before them.”

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	3
1.3	Problem Statement	3
1.4	Downstream Applications of Text Segmentation	5
1.5	Research Questions	6
1.6	Contributions	6
1.7	Document Organization	7
2	State of the Art in Text Segmentation	9
2.1	Literature Review Methodology	9
2.1.1	Identification of the Start Set	10
2.1.2	Screening and Selection Criteria	10
2.1.3	Data Extraction and Iteration Process	11
2.2	Evolution of Text Segmentation Methods	12
2.2.1	Classical and Unsupervised Methods	13
2.2.2	Supervised Neural Architectures	14
2.2.3	LLM-based and Reasoning-oriented Segmentation	15
2.3	Text Segmentation Across Domains	16
2.3.1	Written Documents	16
2.3.2	Dialogue and Spoken Transcripts	17
2.3.3	Segmentation for RAG and Chunking	18
2.4	Evaluation Frameworks	19
2.4.1	Probabilistic and Window-Based Metrics	19
2.4.2	The Edit-Distance Approach: Boundary Similarity (B)	20
2.4.3	Granularity and Window-Tolerant Metrics	20
2.5	Datasets and Benchmarks	21
2.5.1	General-domain Corpora	21
2.5.2	Spoken Dialogue and Meeting Transcripts	21
2.5.3	Domain-specific Corpora	23
2.6	Critical Analysis and Research Gap	23
2.6.1	The Language and Resource Gap	24
2.6.2	Administrative Domain Specificity	24
2.6.3	Challenges in Structuring Municipal Minutes	25
2.7	Summary	25

3	The CitiLink-Minutes Dataset	26
3.1	Municipal Data Challenges, Collection, and Annotation	26
3.2	Dataset Construction and Structural Transformation	29
3.3	Relevance for Text Segmentation	33
3.4	Summary	35
4	SegNSP: A Linear Text Segmentation Framework	38
4.1	Problem Formulation	39
4.2	Experimental Setup	42
4.2.1	Training Setup	44
4.3	Experiments and Results	45
4.3.1	Cross-Domain and Cross-Lingual Results	45
4.3.2	In-Domain Cross-Municipality Generalization	47
4.3.3	Ablation Study	48
4.3.4	Error Analysis	48
4.3.5	Computational Efficiency and Environmental Impact	50
4.4	Downstream Task Evaluation: Summarization	52
4.4.1	Experimental Setup and Baseline Selection	52
4.4.2	Results and Analysis	53
4.5	Summary	54
5	HSeg: A Two-Stage Hierarchical Text Segmentation Framework	56
5.1	Problem Formulation	57
5.2	Proposed Approach: The HSeg Pipeline	58
5.3	Experimental Setup and Evaluation Framework	60
5.4	Results	62
5.4.1	Domain-Specific Evaluation (CitiLink-Minutes)	63
5.4.2	Open-Domain Evaluation (WIKI-727K)	64
5.4.3	Semantic Schema as a Reasoning Scaffold	67
5.5	Discussion	68
5.6	Summary	69
6	Practical Applications and Demonstrations	70
6.1	Interactive Linear Segmentation: The SegNSP Demo	70
6.2	Qualitative Analysis and Model Comparison: HSeg Evaluation Interface	72
6.3	Production Deployment: The CitiLink Platform	73
6.4	Summary	77
7	Conclusions and Future Work	78
	References	82
A	Supplementary Evaluation Results	93
B	HSeg Prompt Templates	94

List of Figures

1.1	The CitiLink research pipeline, illustrating the modular stages of information extraction from municipal minutes.	2
1.2	A conceptual overview of the Text Segmentation task, showing the transformation from a document to a sequence of topically coherent segments.	4
1.3	A dendrogram illustrating the hierarchical multi-level organization of city council minutes, showing macro-level agenda items branching into micro-level specific subjects.	4
3.1	Primary Natural Language Processing (NLP) challenges in municipal documents.	27
3.2	Overview of the four-layer annotation framework implemented in the CitiLink-Minutes dataset (adapted from Fig. 1 in [12]).	29
3.3	The 6-step automated data pipeline transforming raw INCEpTION annotations into the final subject-centric JSON hierarchy.	31
3.4	Summary of the JSON hierarchy for the CitiLink-Minutes dataset, showcasing the nested encapsulation of agenda items, subjects, sub-subjects, summaries, and voting data (adapted from Fig. 3 in [12]).	34
4.1	Overview of the NSP-based segmentation pipeline.	40
4.2	Adaptation of CitiLink-Minutes for linear text segmentation. Macro boundaries (agenda items) and micro boundaries (subjects) are merged to construct contiguous, non-overlapping segments.	43
4.3	CitiLink-Minutes: bimodal distribution of $P(\text{not_next})$, showing a 1.4% overlap in the ambiguous zone $[0.4, 0.6]$	49
4.4	WikiSection: bimodal distribution of $P(\text{not_next})$, broader overlap (4.7%) in the ambiguous zone $[0.4, 0.6]$	50
4.5	CitiLink-Minutes: error distribution across document positions.	51
4.6	WikiSection: error distribution across document positions.	52
4.7	Over- and under-segmentation ratios across CitiLink-Minutes municipalities. . .	53
5.1	Illustration of Hierarchical Text Segmentation (HTS). A document D is partitioned into macro-level sections; each section is subsequently partitioned into micro-level subjects.	57
5.2	Overview of the HSeg segmentation pipeline.	59
6.1	The SegNSP text segmentation demonstration interface, showcasing the interactive threshold slider and real-time document segmentation visualization.	71
6.2	Detailed view of the SegNSP output, explicitly rendering the topic boundary detection confidence scores between extracted segments to aid in error analysis. . .	72

6.3	The Per-Document Metric Distribution view, allowing researchers to isolate model performance by specific municipal domains.	73
6.4	The Document Explorer view (Macro-Level), demonstrating the side-by-side structural alignment of predicted Agenda Items against the ground truth.	74
6.5	The Document Explorer view (Micro-Level), detailing the extraction of individual subjects, their semantic theme matching, and boundary IoU scores.	74
6.6	The Run Comparison view, providing a visual benchmark of evaluation metrics across different generative model configurations.	75
6.7	The CitiLink platform's grid view, displaying extracted subjects as highly scannable cards equipped with high-level voting distributions.	76
6.8	The chronological list view in the CitiLink application, enabling seamless navigation between sequentially discussed subjects.	76
6.9	Detailed modal view for an individual extracted subject, featuring the generated summary, exact transcript boundaries, and voting breakdown.	77

List of Tables

2.1	Summary of Inclusion Criteria	11
2.2	Summary of Exclusion Criteria	11
2.3	Comparative Overview of Primary Text Segmentation Datasets and Benchmarks .	24
3.1	Corpus statistics for the CitiLink-Minutes dataset, distributed by municipality. . .	28
3.2	Corpus statistics characterizing the CitiLink-Minutes dataset, distributed by municipality.	35
3.3	The 22-category municipal topic taxonomy used in the CitiLink-Minutes dataset .	36
4.1	Cross-domain and cross-lingual segmentation results across Portuguese and English datasets using standard train/test splits (German subsets are relegated to Appendix A for readability). Higher $B-F_1/B$ and lower P_k/WD indicate better performance. Values are macro-averaged; best in bold.	46
4.2	Leave-one-municipality-out cross-validation results on CitiLink-Minutes. Higher $B-F_1$ and B , and lower P_k and WD indicate better performance. Best values are in bold.	47
4.3	Ablation of loss components on CitiLink-Minutes.	48
4.4	Comparison of downstream summarization performance (scaled 0–100) using Gemini 2.5 Flash. The control group utilizes perfect manual annotations, while the proposed group relies entirely on SegNSP predicted boundaries.	54
5.1	Stage 1 section segmentation results on the CitiLink-Minutes test split. Higher $B-F_1/B$ and lower P_k/WD indicate better performance.	63
5.2	Stage 2 (Subject level) boundary detection results evaluated within the HSeg pipeline framework.	64
5.3	Hierarchical boundary detection results on the WIKI-727K test set (10% random sample for HSeg models).	65
5.4	Comparison of HSeg performance across dataset scales (WIKI-50 vs. WIKI-727K). Results indicate stable generalization capabilities irrespective of inference volume.	66
5.5	Ablation study isolating the impact of semantic metadata generation on micro-level boundary detection accuracy (CitiLink-Minutes test split, Claude Opus 4.7).	67
A.1	Cross-domain segmentation results for the German WikiSection subsets. Higher $B-F_1/B$ and lower P_k/WD indicate better performance. Values are macro-averaged; best in bold.	93

List of Acronyms

AMALIA	Assistente Multimodal Automático de Linguagem com Inteligência Artificial
IR	Information Retrieval
HTS	Hierarchical Text Segmentation
LLM	Large Language Model
LTS	Linear Text Segmentation
NLP	Natural Language Processing
NSP	Next Sentence Prediction
PII	Personally Identifiable Information
RAG	Retrieval-Augmented Generation
SotA	State-of-the-Art
SDG	Sustainable Development Goal
TS	Text Segmentation

Chapter 1

Introduction

Driven by legal requirements for transparency and the increasing availability of public interest data, municipal public records have grown into vast digital repositories, sparking a growing interest in their computational processing as they become highly attractive resources for a variety of civic and analytical purposes. They serve as the primary legal and historical record of local governance. Nevertheless, navigating these records remains constrained by their often dense and semi-structured nature. Among these, city council minutes represent a critical, but challenging source of information. However, their extensive length, specialized procedural jargon, and lack of explicit thematic markers make automated systematic analysis and information discovery highly challenging. This dissertation tries to address this bottleneck by exploring the task of Text Segmentation (TS), specifically the detection of topic boundaries, within the domain of city council minutes. By evaluating the efficiency of supervised neural architectures and the emerging reasoning capabilities of Large Language Models (LLMs), this work automates the structural decomposition of these documents, thereby enhancing their accessibility and utility for both public officials and citizens.

The remainder of this chapter is structured as follows: Section 1.1 establishes the environment in which this work was developed, specifically detailing its integration within the CitiLink project. Section 1.2 presents the motivation driving this research, followed by the formal problem statement and domain challenges in Section 1.3. Section 1.4 discusses the downstream applications enabled by text segmentation. Finally, the guiding research questions and the main contributions of this dissertation are detailed in Sections 1.5 and 1.6, respectively.

1.1 Context

City council minutes are detailed records of municipal meetings, containing discussions, deliberations, and decisions on a wide range of public topics. These minutes play a critical role in promoting transparency and accountability in local governance, yet their complexity and length pose challenges for both citizens and public servants who seek to access relevant information efficiently.

To tackle the semantic structuring of these texts, this dissertation focuses on developing automated segmentation models. This work is integrated into the CitiLink [77] project¹, a research initiative dedicated to developing Natural Language Processing (NLP) models specifically for European Portuguese. Overall, the project aims to transform the dense text of municipal council meeting minutes into structured, concise, and actionable information, using data sourced from six Portuguese municipalities: Alandroal, Campo Maior, Covilhã, Fundão, Porto, and Guimarães. The overall architectural pipeline of CitiLink, illustrated in Figure 1.1, consists of eight modular stages: (1) detection of boundaries for opening, body, and closing segments; (2) identification of meeting metadata and participant information; (3) **text segmentation into subjects of discussion**; (4) anonymization of personal data; (5) summarization of discussion subjects; (6) recording of individual and party votes; (7) extraction and classification of subjects by municipal department; and (8) the generation of thematic labels. While the project encompasses the full breadth of administrative document processing, this dissertation focuses specifically on stage (3). By identifying precise topic boundaries through the Text Segmentation (TS) task, this work provides the foundational semantic partitioning that anchors the CitiLink architecture. This establishes the coherent thematic units necessary for downstream tasks, ranging from automated summarization and the recording of legislative votes to the systematic classification of municipal themes.

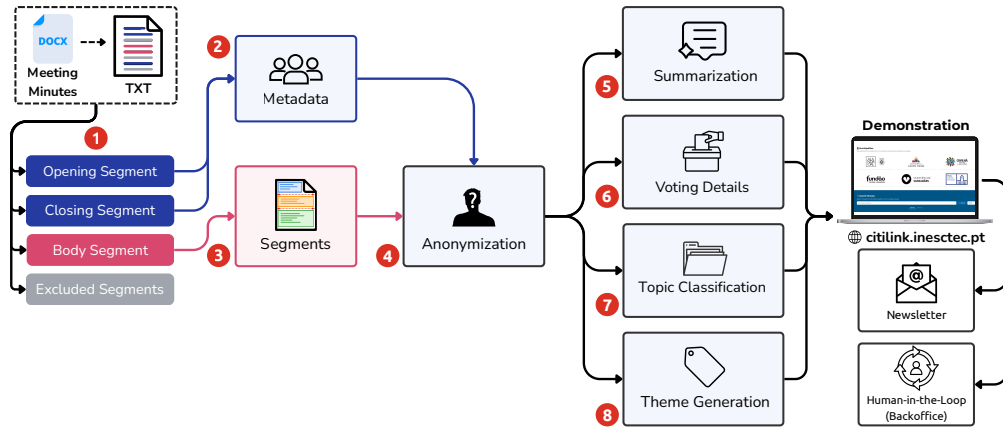


Figure 1.1: The CitiLink research pipeline, illustrating the modular stages of information extraction from municipal minutes.

Automatic Text Segmentation (TS), a task within the broader field of NLP, aims to divide a document into topically coherent sections that correspond to distinct discussion units [29, 40]. While Text Segmentation (TS) has been studied for decades, most research has focused on domains such as news articles (e.g., broadcast transcripts or financial news) or scientific papers (e.g., academic journals and conference proceedings), where discourse structure is clearer and topic transitions are more explicit. In contrast, city council minutes are formal written records of deliberative sessions, often summarizing complex discussions, motions, and administrative decisions in

¹<https://projeto.citolink.inesctec.pt/>

continuous prose. Despite their semi-structured nature, topic boundaries are not explicitly marked, and discussions on related subjects may be interleaved, making automatic segmentation challenging [36].

Recent advances in transformer-based language models, such as BERT and its multilingual or domain-adapted variants [19, 21], have shown promise for long-text processing tasks. However, applying these models to the segmentation of real-world minutes requires addressing issues like data scarcity, domain specificity, and the ability to generalize across languages and municipalities.

1.2 Motivation

The motivation for this work stems from the increasing need to improve access to information in the public administration domain. Municipal minutes contain valuable insights into policy decisions, urban planning, and local governance processes, yet their complexity and length hinder efficient exploration and analysis. By segmenting these documents into meaningful thematic units, it became possible to enable advanced downstream applications such as automatic summarization [15], question answering [68], and information retrieval [80], thus supporting transparency and promoting citizen engagement through easier access to relevant information.

From a research standpoint, the domain also posed a unique opportunity to evaluate the robustness and transferability of segmentation models across different contexts. While large-scale pre-trained models have achieved strong results in benchmark datasets, their behavior in domain-specific scenarios remains underexplored [56]. Developing and evaluating segmentation methods for city council minutes therefore contributed to advancing research in Text Segmentation (TS) and discourse analysis, while simultaneously improving information accessibility in the context of public sector documentation.

1.3 Problem Statement

The task of Text Segmentation (TS) refers to the process of dividing a document into consecutive, topically coherent segments that correspond to distinct discussion topics [40]. In the context of city council minutes, segmentation boundaries typically mark transitions between different sections of the meeting record, such as the discussion of specific proposals, deliberations, or administrative matters.

More concretely, given a document composed of a sequence of sentences or utterances, the goal is to identify where the topic changes, as illustrated in Figure 1.2. Each segment groups together consecutive sentences discussing a common subject, enabling downstream systems to treat each segment as an independent unit for retrieval or summarization.

Unlike domains such as news or scientific writing, municipal minutes present challenges related to their length, domain-specific vocabulary [24], and complex discourse structure [67]. Although they follow an organized narrative flow, topic transitions are often subtle or implicit, requiring models that can capture fine-grained shifts in thematic coherence and discourse context.

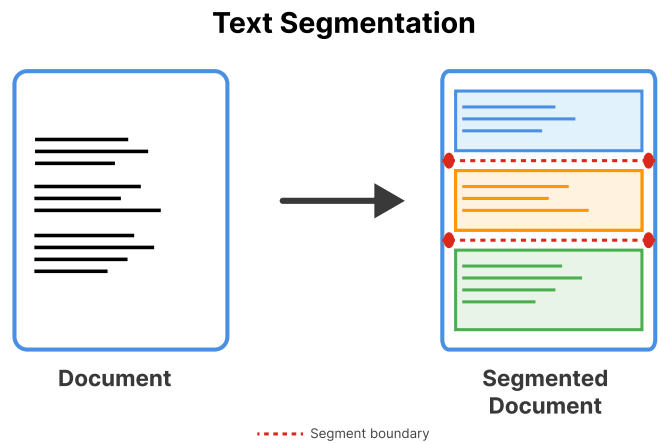


Figure 1.2: A conceptual overview of the Text Segmentation task, showing the transformation from a document to a sequence of topically coherent segments.

To address these shifts, the task is typically framed within two structural paradigms: linear and hierarchical segmentation [9].

Linear Text Segmentation (**LTS**) treats a document as a sequential chain of non-overlapping thematic units [33]. However, this sequential assumption may oversimplify the internal structure of semi-structured records. Hierarchical Text Segmentation (**HTS**) acknowledges that discourse is often nested, where broad "parent" topics encompass multiple specific "child" sub-topics [8]. Instead of a flat list, **HTS** produces a tree-like structure (or dendrogram), as illustrated in Figure 1.3, that reflects a document’s multi-level organization. This multi-level perspective is critical for municipal minutes, where a macro-level agenda item (e.g., "Urban Planning") may contain several distinct but related micro-level discussions (e.g., "Zoning Permits" and "Public Park Maintenance").

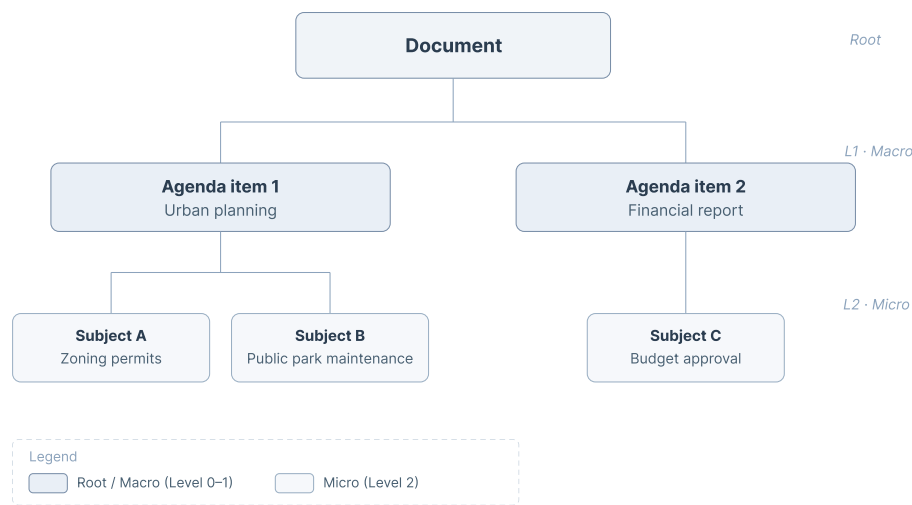


Figure 1.3: A dendrogram illustrating the hierarchical multi-level organization of city council minutes, showing macro-level agenda items branching into micro-level specific subjects.

Therefore, the core problem we are addressing in this dissertation is the accurate and automated detection of both linear and hierarchical topic boundaries in dense, semi-structured municipal records, enabling their transformation into navigable and semantically coherent units.

1.4 Downstream Applications of Text Segmentation

As highlighted in the motivation of this work, document segmentation is not an end in itself, but rather a means to an end for higher-level semantic tasks. The strategic importance of effective segmentation cannot be overstated, as it serves as a critical preprocessing step for a wide range of downstream **NLP** applications.

In the field of Information Retrieval (**IR**), Tiedemann and Mur [80] demonstrate that segmenting documents into smaller, coherent units facilitates more precise passage-level indexing, which significantly improves the contextual relevance of search results compared to document-level retrieval. Similarly, in question-answering systems, the ability to identify topical boundaries is essential for pinpointing precise answers within extensive documents, thereby reducing the noise that often plagues large-scale information extraction [68].

In contemporary Natural Language Processing (**NLP**), these tasks, alongside automatic summarization [15], are increasingly unified under the framework of Retrieval-Augmented Generation (**RAG**) [51]. **RAG** represents a paradigm shift in how information is accessed, combining the strengths of dense information retrieval with the generative capabilities of **LLMs**. In a standard **RAG** pipeline, a document is divided into "chunks" that are indexed in a vector database [82]. When a user submits a query, whether to summarize a specific debate or to answer a question about a local ordinance, the system retrieves the most relevant chunks to serve as the factual context for the model's response [45].

However, the effectiveness of this retrieval step depends heavily on the location of the chunk boundaries. Traditional systems often rely on fixed-window chunking, which divides text into uniform segments based on a predetermined number of characters or tokens [45]. This "semantic-agnostic" approach frequently fractures coherent arguments, separating important context from its associated facts, which can lead to inaccurate or incomplete retrieval [82]. This fragmentation is a primary source of hallucinations in **LLMs**, as the model is forced to reason over incomplete or irrelevant snippets of information [30].

Recent research suggests that topic-aware segmentation, which aligns chunk boundaries with natural thematic shifts, significantly improves retrieval precision [82] and the factual grounding of generated responses [73]. By ensuring that each chunk is a "topically complete" unit, models can capture the full semantic essence of a discussion without the noise of adjacent, unrelated topics [65]. In the context of city council minutes, where a single deliberation may span several pages with subtle sub-topics, transitioning from fixed-window strategies to semantic-aware partitioning is crucial for building reliable search and reasoning tools for legislative archives.

1.5 Research Questions

To guide this research, the following questions are defined, focusing on the efficacy and transferability of modern architectures within the specific domain of administrative records:

- RQ1:** How do Transformer-based models generalize across languages and domains when performing topic boundary detection in semi-structured deliberative documents such as municipal minutes?
- RQ2:** To what extent does the use of a Next Sentence Prediction (NSP) objective effectively capture local discourse continuity for the specific task of identifying topical transitions?
- RQ3:** How do generative Large Language Models (LLMs) compare to foundational unsupervised methods and supervised Transformer-based models in detecting topic boundaries?
- RQ4:** Does the incorporation of hierarchical structural information significantly improve segmentation performance and boundary precision within city council minutes, compared to purely linear modeling approaches?

1.6 Contributions

Building on the research questions outlined above, this dissertation makes the following primary contributions to the field of administrative text processing:

- **Construction of the CitiLink-Minutes Dataset:** To address the scarcity of domain-specific, hierarchically annotated corpora, we contributed to the construction of the CitiLink-Minutes dataset² [12]. This work was published as a peer-reviewed resource paper at ECIR 2026. Comprising 120 manually annotated city council minutes from six Portuguese municipalities, this corpus serves as the fundamental evaluation testbed for this research. Its multi-municipality distribution enables the cross-domain generalization analysis required to answer RQ1, while its nested annotation schema provides the essential ground truth needed to evaluate hierarchical structures against linear baselines (RQ4).
- **SegNSP Model (Linear Text Segmentation):** We introduced a label-agnostic Next Sentence Prediction (NSP) architecture adapted for administrative text. By utilizing a segmentation-aware loss function and domain-balanced sampling, SegNSP established a competitive linear baseline for European Portuguese municipal records ($B-F_1$ of 79.22), outperforming standard general-purpose architectures. This work also forms the basis of a manuscript currently under review [43].
- **HSeg Framework (Hierarchical Text Segmentation):** To address the limitations of linear segmentation in documents with hierarchical structure, we developed HSeg, a configurable, recursive generative LLM pipeline. The method uses sentence-ID grounding and

²<https://github.com/INESCTEC/citilink-dataset>

semantic metadata to guide the extraction of macro- and micro-level discourse segments. On CitiLink-Minutes, HSeg outperforms the SegNSP baseline on micro-level segmentation, reaching a B- F_1 of 89.16. Furthermore, it sets a new state-of-the-art on the large-scale, open-domain WIKI-727K benchmark: using Gemini 3.1 Flash Lite, it improves over the previous CRF-based architecture by 9.06 F_1 points (80.66 vs. 71.60), while frontier models such as Claude Opus 4.7 (78.83 F_1) concurrently achieve the lowest probabilistic error rates (P_k) and the highest precision among all evaluated configurations.

- **Practical Deployment and Civic Application:** Beyond theoretical and empirical advancements, the segmentation model developed in this dissertation has been deployed in a functional demonstration environment and published as a demonstration paper at ECIR 2026 [77]. Our models currently serve as the core structuring engine for the live CitiLink web application³, a real-world civic technology platform designed to enhance public transparency and access to municipal records. This practical integration successfully demonstrates the viability and scalability of our architectures in downstream software systems, effectively bridging the gap between academic NLP research and applied engineering.
- **Open-Source Research Artifacts:** To ensure reproducibility and foster future research, we publicly released the complete source code for both the SegNSP⁴ and HSeg⁵ frameworks on GitHub. Additionally, the fine-tuned SegNSP model has been made available via Hugging Face⁶. To facilitate qualitative evaluation, we also developed and deployed live interactive demonstration platforms for both our SegNSP⁷ and HSeg⁸ frameworks.

1.7 Document Organization

This document is organized into seven chapters, with the current one being Chapter 1. The remaining chapters are structured as follows.

Chapter 2, *State of the Art*, presents the fundamental concepts underpinning this dissertation, covering the evolution of Text Segmentation (TS) approaches from classical unsupervised methods to neural and generative architectures, the role of the NSP objective in discourse modeling, and the specific challenges posed by administrative and municipal documents in the context of NLP.

Chapter 3, *CitiLink-Minutes Dataset*, introduces the first European Portuguese corpus for administrative text segmentation, developed as part of this research. It details the data collection process across six Portuguese municipalities, the pipeline developed to convert flat relational annotations into a standardized hierarchical schema, and the statistical characterization of the resulting resource, which comprises 120 documents and 2,885 manually annotated subjects of discussion.

³<https://citilink.inesctec.pt/>

⁴<https://github.com/jmisidro/segnsp>

⁵<https://github.com/jmisidro/hseg>

⁶<https://huggingface.co/inesctec/citilink-nsp-segmentation-pt>

⁷https://huggingface.co/spaces/inesctec/citilink_nsp_segmentation-demo

⁸https://huggingface.co/spaces/jmisidro/hseg_demo

Chapter 4, *SegNSP*, presents the first of two segmentation frameworks developed in this dissertation. It formulates **LTS** as a boundary-sensitive coherence classification problem, describes the architectural adaptations to the **NSP** objective, including domain-balanced sampling, harder negative construction, and a segmentation-aware loss function, and reports results on both the CitiLink-Minutes corpus [12] and the cross-lingual WikiSection benchmark [3].

Chapter 5, *HSeg*, presents the hierarchical segmentation framework, which addresses the structural limitations of linear approaches by employing a recursive, **LLM**-based pipeline to recover nested discourse structures at both the macro agenda-item and micro subject-of-discussion levels. It further reports a comprehensive multi-model evaluation on CitiLink-Minutes [12] and validates the framework’s cross-domain generalizability on the large-scale WIKI-727K benchmark [47], including a pilot evaluation on a WIKI-50 subset to validate sampling robustness, where new state-of-the-art results are established.

Chapter 6, *Practical Applications and Demonstrations*, describes the integration of the developed frameworks into the CitiLink civic technology platform and the associated demonstration tools, illustrating the real-world applicability of the segmentation models to European Portuguese municipal archives.

Finally, Chapter 7, *Conclusions and Future Work*, synthesizes the outcomes of the conducted research, revisits the four research questions guiding this dissertation, and outlines the most promising directions for future investigation, including segmentation-aware **RAG**, the fine-tuning of open-weight models on administrative text, and the exploitation of the sub-subject annotation layer introduced in CitiLink-Minutes v1.1.0.

Chapter 2

State of the Art in Text Segmentation

This chapter establishes the theoretical foundation for Text Segmentation, with a particular focus on detecting topic boundaries in city council minutes. To ensure a comprehensive and unbiased analysis, the literature review process is first detailed, specifically the snowballing methodology employed to gather and select the works discussed throughout this chapter. Following this methodological overview, the chapter documents the evolution of Text Segmentation models, beginning with foundational unsupervised methods that rely on lexical statistics, and progressing towards modern supervised neural architectures that learn segmentation cues directly from annotated data. Furthermore, it provides a comparative analysis of evaluation frameworks and contextualizes the classic task of Text Segmentation within the contemporary challenge of document chunking for Retrieval-Augmented Generation (RAG) systems.

The remainder of this chapter is structured as follows: Section 2.1 details the snowballing methodology used to identify relevant literature. Section 2.2 documents the evolution of text segmentation methods, spanning classical, neural, and LLM-based approaches. Section 2.3 explores how these methods are applied across different contexts, including written documents, dialogue, and RAG chunking. Section 2.4 discusses the evaluation metrics used in the field, and Section 2.5 outlines the primary datasets and benchmarks. Finally, Section 2.6 presents a critical analysis of the literature, identifying the research gaps this dissertation aims to address.

2.1 Literature Review Methodology

This dissertation employs an exploratory literature review to establish a theoretical foundation for Text Segmentation, with a particular focus on detecting topic boundaries in city council minutes. Rather than relying solely on database-driven search strings, which may miss relevant studies due to inconsistent terminology in this evolving field, this study adopts a snowballing sampling technique [84]. Snowballing is the iterative process of using the reference lists of a paper (backward snowballing) or the citations of a paper (forward snowballing) to identify additional papers. This

methodology is particularly effective for identifying sources in specific domains, where protocol-driven search approaches might be insufficient. The following sections detail the procedural steps taken, ensuring the reproducibility and reliability of the review process.

2.1.1 Identification of the Start Set

The review follows the snowballing guidelines for systematic literature studies, as proposed by Wohlin [84]. The procedure is iterative, beginning with the identification of a "Start Set" of papers, followed by multiple rounds of backward and forward searching until no new relevant papers are found.

The first and most critical step in the snowballing procedure is the identification of a "Start Set" (or Seed Set) of papers. According to Wohlin [84], a good start set should cover different communities, publishers, and years to avoid bias and ensure diversity. For this study, the start set was selected to cover the fundamental evolution of the field: from unsupervised lexical cohesion methods, to dialogue-specific segmentation, to the current State-of-the-Art (SotA) neural approaches, and finally to domain-specific datasets.

The Start Set comprises four anchor papers and was designed to cover the evolution of topic segmentation research, beginning with the foundational lexical-cohesion paradigm and progressing toward domain-relevant multi-party dialogue settings and modern neural approaches. We first anchored the review with Hearst [40], who introduced *TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*, a seminal baseline that operationalizes segmentation via lexical co-occurrence patterns. Building on this, we included Galley et al. [29] (*Discourse Segmentation of Multi-Party Conversation*) to capture the shift from clean expository text to conversational data, introducing LCSeg and incorporating dialogue cues such as speaker changes and silence. To represent the supervised neural paradigm enabled by large-scale resources, we selected Koshorek et al. [47] (*Text Segmentation as a Supervised Learning Task*), which frames segmentation as a supervised problem and introduces the WIKI-727K dataset. Finally, to anchor the review in meeting-style administrative discourse, we included Hu et al. [41] (*MeetingBank: A Benchmark Dataset for Meeting Summarization*), which, although focused on summarization, provides a closely related domain context and helps ensure the snowballing process captures recent advances relevant to municipal minutes.

2.1.2 Screening and Selection Criteria

Following the establishment of the Start Set, the iterative process of backward and forward snowballing was initiated. To ensure consistency, explicit inclusion and exclusion criteria were applied during the screening phases.

The inclusion criteria, detailed in Table 2.1, were designed to capture the three primary pillars of this dissertation: the evolution of segmentation algorithms, the specific challenges of multi-party dialogue, and the applicability of these methods to the Portuguese language.

Table 2.1: Summary of Inclusion Criteria

Category	Selection Rationale
<i>Topic Relevance</i>	Studies proposing novel algorithms for linear or discourse segmentation, or creating datasets specifically designed for evaluating segmentation tasks.
<i>Domain Relevance</i>	Research centered on multi-party dialogue, meeting transcripts, or parliamentary/legislative proceedings (similar to the city council context).
<i>Language Specificity</i>	Works focusing on European Portuguese or developing cross-lingual/language-independent models that can be effectively applied to Portuguese text.
<i>Methodological Rigor</i>	High-quality, peer-reviewed contributions published in reputable international journals or core NLP/AI conferences (e.g., ACL, ECIR, EMNLP, LREC).

While the inclusion criteria defined the scope, the exclusion criteria (see Table 2.2) were applied to prune the search results and maintain a sharp focus on textual analysis. These boundaries are particularly important in the field of meeting analysis, where research often overlaps with the fields of signal processing and computer vision, both of which fall outside the scope of this study.

Table 2.2: Summary of Exclusion Criteria

Category	Reason for Exclusion
<i>Modality</i>	Studies focusing exclusively on acoustic signals (e.g., pitch, energy) or visual cues (e.g., camera angles, gesture recognition) without a textual/NLP component.
<i>Granularity</i>	Research restricted to micro-level tasks such as sentence boundary detection or punctuation restoration without addressing macro-level topic shifts.
<i>Incomplete Results</i>	Short papers, posters, or extended abstracts that do not provide a detailed description of the methodology or formal evaluation metrics.
<i>Redundancy</i>	Preliminary versions of studies (e.g., conference papers) that have been superseded by more comprehensive, peer-reviewed journal articles by the same authors.

2.1.3 Data Extraction and Iteration Process

The data extraction phase followed the systematic, iterative steps outlined by Wohlin [84], ensuring a comprehensive search of both the historical foundations and the contemporary frontiers of the field. The process was divided into two search directions: backward and forward snowballing.

During the backward snowballing phase, the reference list of each paper in the current set was examined in detail. First, titles were screened for relevance to topic segmentation or Portuguese linguistics. Then, the publication venue and authors were scrutinized for credibility and fit. Papers that passed this initial filter underwent a contextual review, which involved examining the specific section in the source text where the paper was cited. This allowed for an assessment of the paper’s potential contribution before proceeding to a full-text review. Finally, the abstract and complete manuscript were evaluated against the inclusion criteria to determine whether the study should be added to the seed set for the next iteration.

In parallel, forward snowballing was conducted by identifying more recent work citing the papers in the current set. The abstracts of these papers were examined using citation databases to identify newer methodologies, datasets, or critical perspectives relevant to the segmentation of city council minutes. This iterative cycle of backward and forward searches continued until the termination criterion was met, which occurred when a full iteration yielded no new papers that met the defined inclusion criteria.

This approach ensures that the review captures the evolution of the field, bridging the gap between foundational methodologies and contemporary **SotA** advancements. By adhering to this rigorous process, the resulting literature review identifies and synthesizes the critical intersections most pertinent to the research objectives. Ultimately, this provides a robust theoretical and empirical foundation for the subsequent chapters, ensuring that the study is positioned within the broader academic discourse while remaining focused on its specific domain of application.

2.2 Evolution of Text Segmentation Methods

Having established the literature review methodology, it is essential to formalize the computational concepts that define this field before detailing the chronological evolution of specific algorithms. Text Segmentation, frequently referred to as topic segmentation or discourse segmentation, is a fundamental task in Natural Language Processing (**NLP**) focused on partitioning continuous text into topically coherent units. At its core, the problem is framed as a boundary detection task, where the objective is to identify the specific points at which a latent semantic shift occurs, be it at paragraph, sentence or utterance level [40, 47]. In the literature, the resulting "segments" are defined not merely as arbitrary textual chunks, but as internally cohesive blocks [72]. Early computational models framed these segments as contiguous spans of text bound by high lexical overlap [16, 40], whereas modern neural approaches define them as dense contextual clusters within a learned representation space [36].

Computationally, the task is typically formulated within two distinct structural paradigms: linear and hierarchical segmentation. Linear Text Segmentation models a document as a flat, one-dimensional sequence of non-overlapping thematic chunks [33]. Under this paradigm, segmentation is traditionally cast as a binary classification or sequence-labeling problem, where each candidate boundary is evaluated independently to determine if a transition has occurred [47].

However, this sequential assumption often oversimplifies complex discourse. Hierarchical Text Segmentation (**HTS**) acknowledges that texts frequently exhibit nested, multi-level structures, where broad macro-topics encompass several specialized sub-topics [23]. Rather than producing a flat list, **HTS** algorithms are typically formulated as parsing or divisive clustering problems that generate a tree-like hierarchy of the document's semantic structure [3]. This requires models to evaluate the relative strength of semantic shifts to differentiate between a major global transition and a minor local digression.

The methodologies employed to solve both linear and hierarchical boundary detection have undergone a significant transformation over the past decades. The following subsections trace this evolution, moving from early unsupervised heuristic algorithms to the adoption of advanced neural and language-model-based architectures.

2.2.1 Classical and Unsupervised Methods

Early research in Text Segmentation was predominantly unsupervised, driven by the need to identify structure in large corpora without the prohibitive cost of manual annotation. These models typically rely on the linguistic theory of lexical cohesion [39, 61], which states that related vocabulary tends to co-occur within a single topical segment, and that boundaries can be inferred from drops in lexical overlap.

The **TextTiling** algorithm [40] remains the most influential foundational model in the field. It operates on the premise that thematic shifts are marked by significant changes in vocabulary. The algorithm proceeds through three distinct stages: tokenization into fixed-size "token sequences" (to remain robust against missing sentence boundaries), similarity determination via cosine similarity between adjacent blocks, and boundary identification. The latter is achieved by analyzing a coherence graph where "valleys" (local minima) represent potential topic shifts. Hearst [40] introduced the concept of a "depth score" to measure the significance of these valleys relative to the surrounding "peaks," ensuring that only the most abrupt semantic shifts are labeled as boundaries. Despite its pioneering status, **TextTiling** is highly sensitive to lexical variety and word choice, often struggling with short-form discourse or texts where synonyms are used frequently [29].

To address the "vocabulary mismatch" problem, where different words refer to the same underlying topic, researchers shifted toward probabilistic frameworks. Beeferman, Berger, and Lafferty [10] utilized adaptive language models that identified "trigger pairs," which are words that significantly increase the probability of a topic shift when they appear in a new context. A major advancement was **TopicTiling** [75], which evolved the window-based approach of Hearst [40] by integrating **Latent Dirichlet Allocation**. By mapping words to a latent topic space, **TopicTiling** calculates coherence based on the distribution of inferred topic IDs rather than raw word counts. This provides a higher level of semantic abstraction, allowing the model to recognize a topic shift even when the specific vocabulary remains diverse. The most recent innovation in this category is the **Unsupervised Mutual Learning Framework** proposed by Xu et al. [87]. Recognizing that topic structure and rhetorical structure are deeply intertwined, this model utilizes two linguistically

grounded hypotheses: *Local Discourse Coupling*, where rhetorical cues (e.g., contrast or elaboration) enhance topic-aware information flow, and the *Global Topology Constraint*, where overall topic patterns constrain the distribution of rhetorical relations. By allowing these two tasks to mutually reinforce each other, the model achieves **SotA** unsupervised performance without requiring any task-specific manual annotations.

A different structural perspective is offered by graph-based models, which move away from linear windows. The **C99** algorithm [16] represents a document as a self-similarity matrix of sentences and applies a ranking transform to make the similarity values robust to local noise. By recursively partitioning this matrix, C99 identifies segments that maximize internal similarity. Building on the power of word embeddings, **GraphSeg** [36] constructs a semantic relatedness graph where each node represents a sentence and edges represent semantic similarity. Instead of finding valleys in a sequence, GraphSeg identifies topical segments by finding "maximal cliques", which are sub-graphs where all sentences are highly connected to one another. This approach captures non-linear dependencies and provides a denser semantic representation than earlier sparse-vector methods. In the specialized domain of dialogue, Xing and Carenini [86] introduced **Utterance-Pair Coherence Scoring**. This method leverages the supervisory signals from a BERT-based model to measure the topical relevance between adjacent utterances. While the coherence model is trained on synthetic data (e.g., distinguishing real utterance pairs from random ones), it is applied in an unsupervised manner to identify boundaries in conversational transcripts, effectively bridging the gap between deep neural representations and unsupervised inference.

2.2.2 Supervised Neural Architectures

The emergence of large-scale, annotated datasets, most notably the **WIKI-727K** corpus introduced by Koshorek et al. [47], marked a paradigm shift in the field. By providing over 700,000 Wikipedia documents with clear hierarchical structures, these resources allowed researchers to reframe Text Segmentation as a supervised sequence labeling problem. This shift enabled the development of deep neural architectures capable of learning complex, non-linear signals of topical change directly from raw data.

Initial neural approaches leveraged the sequential modeling strengths of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). The **TextSeg** (Bi-LSTM) model by Koshorek et al. [47] remains a benchmark in this category, employing a hierarchical Bidirectional LSTM. This architecture operates in two stages: a lower-level encoder generates vector representations for individual sentences, which are then fed into a higher-level document-level LSTM to predict binary boundary labels. Similarly, Badjatiya et al. [6] integrated attention mechanisms with CNNs to identify salient features within sentence windows, demonstrating that local feature extraction could effectively signal a thematic shift.

The introduction of the Transformer architecture revolutionized Text Segmentation by allowing for more nuanced modeling of both global and local context, effectively capturing long-range dependencies that were previously difficult for RNNs to maintain. Lukasik et al. [56] pioneered the application of pre-trained encoders with **Cross-Segment BERT**. Their work demonstrated

that by feeding a local window of text, namely the sentences immediately surrounding a potential boundary, into a classifier, a model could achieve high performance without the computational overhead of processing the entire document. This highlighted that local semantic shifts are often the most salient indicators of topical change.

A pivotal advancement in this era was the development of the **SECTOR** framework by Arnold et al. [3]. Unlike previous models that focused solely on boundary detection, SECTOR introduced latent topic embeddings learned through a bidirectional LSTM architecture enhanced by Bloom filter embeddings. This allowed the model to simultaneously segment a document and perform topic classification for each section. Perhaps the most lasting contribution of this work was the introduction of the **WikiSection** [3] dataset. Prior to its release, researchers relied heavily on the Wiki-727K [47] dataset, which, while large, was often criticized for its lack of domain diversity and inconsistent labeling. WikiSection provided a more robust and specialized benchmark, consisting of approximately 38k Wikipedia articles structured into hierarchical sections across diverse domains such as "Cities" and "Diseases." Its importance stems from its clean, structured nature, which includes both English and German portions, making it the most relevant and widely used dataset for evaluating a model's ability to handle domain-specific vocabulary and hierarchical topical transitions. By providing a framework where labels represent standardized section headings (e.g., "History," "Demographics," "Geography"), it enables models to learn the structural conventions of specific genres rather than just detecting lexical shifts.

Building upon the success of hierarchical modeling, Lo et al. [55] introduced **Transformer²**, a "transformer over transformer" framework. This architecture utilizes a bottom-level transformer to encode sentence-level semantics, which are then fed into a second, upper-level transformer that models the document-level discourse. To ensure high-quality topic coherence, the authors implemented a multi-task loss function that requires the model to simultaneously recover segment boundaries and predict topic labels. This approach rests on the hypothesis that knowing what a topic is significantly improves the model's ability to identify where that topic naturally concludes, leading to more semantically consistent segments.

To address the inherent noise and computational complexity of long-form documents, Bai et al. [7] developed **SegFormer**. This model introduces a controllable range of attention, using unidirectional attention blocks to prevent the model from being distracted by distant, irrelevant information, which is a common issue in standard self-attention mechanisms. SegFormer employs a novel context aggregator and an iterative prediction algorithm that progressively refines boundary locations. This iterative approach allows the model to "stress-test" candidate boundaries against surrounding context, significantly mitigating the interference caused by the informal or redundant language often found in extensive transcripts.

2.2.3 LLM-based and Reasoning-oriented Segmentation

Parallel to clustering and sequence labeling, the use of Large Language Models has introduced "deductive" capabilities to the task. Lee et al. [49] proposed **Def-DTS** (Deductive Topic Segmentation), which leverages the reasoning power of LLMs to identify topic shifts. Instead of purely

statistical similarity, Def-DTS prompts the model to perform a multi-step analysis: first identifying the core intent of an utterance and then deductively determining if that intent aligns with the preceding segment. This shift from "similarity-based" to "reasoning-based" segmentation has shown significant promise in handling open-domain dialogues where thematic shifts are subtle.

Beyond dynamic chunking, recent research has increasingly focused on utilizing Large Language Models (LLMs) to explicitly recover and model multi-level, hierarchical document structures. For instance, in the domain of complex educational and legal texts, frameworks like HiPS (Hierarchical PDF Segmentation) have demonstrated that LLMs can successfully refine boundary detection to extract deeply nested segmentations without relying on explicit structural markers [83]. Similarly, to bridge the gap between linear segmentation and hierarchical information retrieval, recent approaches have proposed integrating supervised linear segmentation with unsupervised LLM-driven clustering to build multi-level semantic chunks that capture broader discourse relationships [64]. Furthermore, frameworks such as LATTICE explore organizing document contexts into semantic trees via top-down divisive and bottom-up agglomerative summarization strategies using LLMs [38]. This shift toward LLM-guided, multi-level structural parsing provides a critical conceptual foundation for the generative hierarchical extraction required to navigate highly structured administrative records.

2.3 Text Segmentation Across Domains

2.3.1 Written Documents

A significant evolution in text segmentation for formal written documents is the shift toward joint tasks that combine structural analysis with content generation. For instance, Liu, Zhu, and Zeng [54] introduced the *SegNews* framework, which performs end-to-end segmentation-based news summarization. By jointly segmenting a news article and generating a corresponding heading-style summary for each section, this approach provides a significantly more digestible way for users to consume long-form content. This method addresses the cognitive load associated with reading long documents by presenting information in thematically consistent, summarized blocks, rather than as a single, overwhelming narrative.

This hierarchical approach has been further specialized for semi-structured documents. For example, Virameteekul [81] proposed a model specifically for book chapter segmentation that utilizes **XLNet** connected to a CNN to extract dense paragraph-level semantics. By designing an attention mechanism that specifically measures the similarity between consecutive paragraphs, this model addresses the ambiguity of language in long-form narratives, significantly outperforming standard BERT-based baselines in identifying chapter transitions.

Beyond these general-domain advancements, supervised Transformer models have demonstrated remarkable utility in highly specialized fields. Aumiller et al. [5] adapted these techniques for the structural segmentation of legal documents, specifically Terms-of-Service. Their system frames the problem as "topical change detection," performing independent classifications across

paragraphs to capture the unique, often rigid, structural peculiarities of legal prose. By focusing on the functional roles of paragraphs within a contract, their work illustrates that Text Segmentation in specialized domains requires models to understand not just what is being said, but how the document's formal structure dictates the transition of information.

2.3.2 Dialogue and Spoken Transcripts

While these applications are highly beneficial for formal written documents, the role of Text Segmentation becomes especially critical when processing transcripts of spoken language. Such texts frequently lack the explicit structural markers, such as paragraphs, headers, or chapters, which are ubiquitous in formal writing, making them inherently difficult to navigate and analyze [72]. Beyond the challenges of informal discourse, modern research has highlighted the necessity for models that are robust to the irregularities of real-world data. Frohmann et al. [28] emphasize that many existing systems rely heavily on standard linguistic features that may be absent in noisy environments. In their *Segment Any Text* approach, they advocate for the development of robust, zero-shot models capable of maintaining high performance even in the face of missing punctuation or varied domain-specific styles, ensuring that segmentation can be applied reliably across a truly diverse range of text sources.

Segmenting spoken transcripts and multi-party dialogues presents unique computational challenges that distinguish the task from standard document segmentation. Unlike written text, transcripts often lack explicit punctuation, contain disfluencies (e.g., "uhm," "ah"), and are characterized by speaker overlaps and rapid topical shifts. As noted by Purver [72], these structural deficiencies necessitate models that can look beyond surface-level syntax to capture the underlying discourse flow.

An early milestone in this domain was the **LCSeg** algorithm developed by Galley et al. [29]. LCSeg is grounded in the concept of *lexical chains*, which are sequences of semantically related words that span across multiple utterances. By calculating a cohesion score based on the density and overlap of these chains, the model identifies segments where specific vocabulary is consistently sustained. A key innovation of LCSeg was its integration of non-lexical conversational cues, such as the duration of silences and the presence of cue phrases (e.g., "okay," "anyway"), which act as strong signals for topic transitions in spontaneous speech.

With the rise of deep learning, researchers began adapting classical frameworks for the neural era. Ghinassi et al. [32] introduced **DeepTiling**, a neural adaptation of the TextTiling algorithm that replaces raw word counts with embeddings from pre-trained language models like RoBERTa. Their analysis revealed a critical finding for the dialogue domain: while these models achieve high performance, they often function as "local cue detectors" rather than global coherence modelers. In domains like news broadcasts or formal meetings, models tend to over-rely on specific introductory phrases or speaker change signals rather than holistically measuring the semantic shift between blocks of text.

A significant challenge in supervised Text Segmentation is the mismatch between training data and real-world transcripts. Ghosh et al. [34] conducted a comprehensive analysis of the generalization capabilities of **SotA** models on unstructured conversational datasets. Their findings indicate that current strategies of pre-training on structured corpora, such as the Wiki-727K dataset, do not transfer effectively to the noisy, informal nature of chat data. To address the inherent label imbalance in dialogue, where non-boundary sentences far outnumber boundary sentences, they proposed the use of Focal Loss. This approach encourages the model to focus on "hard" examples (actual boundaries) and significantly improves segmentation results in semi-structured environments when compared to standard Cross-Entropy loss.

More recent work has explored the multi-level nature of human conversation. Gklezakos, Misiak, and Bishop [35] developed **TreeSeg**, an unsupervised hierarchical model that utilizes divisive clustering on sentence embeddings. TreeSeg is particularly valuable for meeting transcripts as it generates a tree-like structure, allowing users to navigate content at varying levels of granularity, from broad agenda items down to specific sub-discussions.

Finally, bridging the gap toward the Portuguese context, Freisinger et al. [27] investigated the performance of **SotA** models on Portuguese lecture transcripts. Their study demonstrates that while multilingual models provide a baseline for zero-shot segmentation, achieving the precision required for educational or administrative archives requires the integration of domain-specific acoustic and rhythmic features. This research highlights the inherent difficulty of applying models trained on written, English-centric data to the prosodic and structural nuances of spoken Portuguese discourse.

2.3.3 Segmentation for RAG and Chunking

In Retrieval-Augmented Generation (**RAG**) systems, Text Segmentation has gained renewed relevance as a mechanism for producing coherent retrieval units, commonly referred to as chunks. In a standard **RAG** pipeline, the quality of the generated response is fundamentally limited by the relevance of the retrieved context. Traditionally, this was managed through static, rule-based methods, such as splitting text by fixed character counts or paragraph markers. However, these methods often result in "broken" contexts, where a single semantic unit is bisected across two chunks. As noted by Zhao et al. [90], this "weakest link" effect significantly degrades retrieval precision and increases the likelihood of model hallucinations.

To move beyond the limitations of fixed-length windows, recent research has leveraged the reasoning capabilities of **LLMs** to identify optimal, semantic boundaries. Duarte et al. [22] proposed **LumberChunker**, a framework grounded in the principle of "semantic independence." Instead of arbitrary splits, LumberChunker iteratively prompts an **LLM** to examine a group of sequential passages and identify the exact point where the narrative or thematic content begins to diverge. This dynamic approach ensures that each chunk encapsulates a complete, standalone idea, which significantly improves retrieval performance (measured by DCG@20) compared to brute-force context ingestion methods like the long-context Gemini 1.5 Pro [22]. Similarly, Wang et al. [82]

introduced the **PIC (Propositional Information Chunking)** method, which argues that "document segmentation matters" as much as the retrieval algorithm itself. PIC dynamically groups sentences based on their semantic similarity to a central summary of the document, ensuring that each chunk remains relevant to a specific theme. This method has demonstrated substantial gains in open-domain question-answering benchmarks by providing "evidence-dense" chunks that minimize the "lost-in-the-middle" phenomenon common in long-context models [82].

While **LLM**-driven segmentation offers high precision, it is often computationally expensive for large-scale document stores. To address this trade-off, Zhao et al. [90] developed the **Mixture-of-Chunkers (MoC)** framework. MoC introduces two novel, quantifiable metrics for chunking quality: *Boundary Clarity* (the degree to which a boundary aligns with a logical shift) and *Chunk Stickiness* (the internal semantic consistency of a unit). The system utilizes a lightweight, multi-granularity router that activates specialized "meta-chunkers" based on the complexity of the text. Crucially, instead of generating full text, the meta-chunkers output structured regular expressions that identify start and end markers, drastically reducing computational overhead while maintaining the accuracy of larger **LLMs** [90].

A persistent challenge in neural Text Segmentation is the loss of semantic information when long documents are pre-processed into smaller segments due to model input constraints. Ni et al. [66] addressed this with **CrossFormer**, a transformer-based model featuring a Cross-Segment Fusion Module. Unlike standard models that process segments in isolation, CrossFormer explicitly models the latent semantic dependencies between adjacent segments. When integrated into a **RAG** pipeline, CrossFormer replaces traditional rule-based splitters, producing chunks that preserve the inter-segment correlations necessary for answering complex, multi-hop queries that span across different parts of a document [66].

Together, these advancements represent a paradigm shift in how **RAG** systems handle data, moving from passive, structural partitioning toward proactive, semantic "document memory" construction. By ensuring that each retrieved unit is topically coherent and semantically self-contained, these models provide the structured foundation necessary for reliable and accurate **LLM** reasoning.

2.4 Evaluation Frameworks

The evaluation of Text Segmentation is notoriously difficult due to the inherent subjectivity of topic boundaries and the "near-miss" problem, where a model identifies a boundary only one or two sentences away from the ground truth. Traditional classification metrics like Accuracy or standard F1 are often considered inadequate because they treat a near-miss as a total failure, failing to reward models for identifying the general vicinity of a topical shift. To address this, specialized probabilistic and window-based metrics have been developed.

2.4.1 Probabilistic and Window-Based Metrics

The first specialized metric to gain widespread adoption was P_k [10]. It operates by sliding a window of size k (typically set to half the average segment length) across the document and cal-

culating the probability that two sentences chosen at random within that window are incorrectly classified as being in the same or different segments. P_k has been criticized for its "penalty bias": it penalizes false negatives (missing a boundary) more severely than false positives (inserting an extra boundary) and is highly sensitive to the distribution of segment lengths.

To mitigate these biases, Pevzner and Hearst [71] introduced **WindowDiff (WD)**. Unlike P_k , which only checks if two points are in the same segment, WindowDiff compares the count of boundaries within the sliding window between the hypothesis and the reference. If the count differs, a penalty is applied. While WindowDiff is more robust than P_k , it still struggles with the "near-miss" phenomenon and can be difficult to interpret when comparing datasets with vastly different boundary densities.

2.4.2 The Edit-Distance Approach: Boundary Similarity (B)

A notable advancement in evaluation methodology occurred with the work of Fournier [26], who proposed **Boundary Similarity (B)**. This metric is grounded in the **Boundary Edit Distance (BED)**, which calculates the minimum number of operations (additions, deletions, or transpositions) required to transform the predicted segmentation into the ground truth.

The primary innovation of B is its ability to award partial credit for near-misses through the "transposition" operation. If a model places a boundary slightly off-center, it is treated as a transposition rather than a combination of a false positive and a false negative. This makes B far more aligned with human intuition than P_k or WindowDiff. Furthermore, Fournier introduced **B- F_1** [26], which utilizes the BED-based confusion matrix to derive precision and recall. By grounding F_1 in edit distance rather than exact matches, B- F_1 provides a balanced measure that is resilient to the structural noise often found in long-form transcripts.

2.4.3 Granularity and Window-Tolerant Metrics

While Text Segmentation models have grown more sophisticated, a persistent challenge in conversational and hierarchical datasets remains: granularity mismatch. In many conversational datasets, there may be multiple "correct" ways to segment a text depending on whether one focuses on broad agenda items or fine-grained sub-topics.

To address this variance, Coen [17] introduced the **Window-tolerant F1 (W-F1)** and **Boundary Density (BOR)** metrics. W-F1 extends the concept of fuzzy matching by allowing a boundary to be considered a true positive if it falls within a predefined slack window. This acknowledges that exact-sentence alignment is often less critical than identifying the correct transitional zone. Furthermore, when paired with BOR, which measures the frequency of predicted boundaries, these metrics allow for a more nuanced distinction between models that accurately identify topic transitions and those that inflate performance simply by over-segmenting the text. This granularity-aware evaluation is particularly relevant for applications such as **RAG**, where the utility of a text chunk depends on its semantic completeness rather than its strict alignment with a single, rigid ground-truth boundary.

2.5 Datasets and Benchmarks

The development of Text Segmentation models has been closely tied to the availability, scale, and annotation schema of available corpora. Early research relied on small, synthetically constructed datasets, whereas the deep learning era has shifted toward massive, naturally structured web scrapings. This section reviews the primary corpora used to train and evaluate modern segmentation architectures, summarized in Table 2.3.

2.5.1 General-domain Corpora

A major expansion in the scale of supervised Text Segmentation data was achieved by leveraging the inherent structure of Wikipedia, which provides a natural ground truth via its user-edited section headers.

- **WIKI-727K** [47]: Comprising over 727,000 documents, this remains the largest general-purpose dataset in the field. Particularly valuable for hierarchical research, it utilizes the complete, nested table of contents of Wikipedia articles as ground-truth boundaries. When introduced alongside the baseline **TextSeg** (Bi-LSTM) model, it established a benchmark P_k of 0.222 [47], though recent Transformer-based architectures have pushed this error rate below 0.15. While its scale is a massive asset, its automated extraction introduces label noise where section headers reflect formatting choices rather than genuine thematic shifts.
- **WikiSection** [3]: To address the thematic ambiguity of WIKI-727K, WikiSection was introduced as a domain-filtered, linear benchmark. It contains approximately 38k articles in English and German, strictly restricted to two domains: "Cities" and "Diseases". Crucially, it features standardized section labels (e.g., *History*, *Geography*), allowing models to learn predictable, genre-specific discourse flows. Evaluated using the **SECTOR** architecture, the dataset yielded a baseline P_k of 0.115 on the English Cities subset [3]. However, its flat annotation schema omits nested sub-sections, limiting its utility for **HTS** tasks.

2.5.2 Spoken Dialogue and Meeting Transcripts

While captured via acoustic recording, spoken discourse corpora are vital to **NLP** segmentation research when accompanied by human-verified, time-aligned text transcripts. Segmenting these transcripts presents unique computational hurdles absent in written prose, such as conversational disfluencies, overlapping speech, missing syntax, and highly collaborative, negotiated topic transitions [72, 76]. To distinguish between foundational literature and peripheral testing grounds, these resources are categorized into primary meeting/municipal benchmarks and secondary conversational references.

Primary Meeting Benchmarks

- **ICSI Meeting Corpus** [76]: Comprising 75 naturally occurring academic and technical meetings (approx. 72 hours of audio and 800k transcribed words), this corpus serves as a classical pillar for spoken discourse analysis. Because it captures unscripted, highly technical group debates, topic boundaries are notoriously fluid. Historically benchmarked using the lexical-cohesion **LCSeg** algorithm, it yielded a baseline probabilistic error rate (P_k) of 0.318 [29].
- **AMI Meeting Corpus** [13]: Consisting of 100 hours of transcribed scenario-based design meetings, AMI provides a controlled environment for tracking multi-party, goal-oriented topic progression. Modern neural evaluations on this corpus, such as the **DeepTiling** architecture (RoBERTa-based), report WindowDiff scores of approximately 0.284, highlighting the persistent difficulty of resolving multi-speaker thematic drift [32].
- **MeetingBank** [41]: Although curated primarily for summarization, MeetingBank is arguably the most relevant external dataset for this dissertation. It comprises 1,366 sessions from the city councils of six major U.S. municipalities, totaling over 3,500 hours of transcribed proceedings. The dataset is notable for its chunk-level alignment methodology: rather than processing an entire multi-hour session at once, the authors partitioned raw transcripts into localized blocks and mapped them directly to the professionally written agenda items in the official published minutes. For Text Segmentation, these agenda mappings provide an exceptionally high-quality ground truth for legislative topic shifts. While monolingual (English), it serves as the definitive high-resource reference model for procedural municipal discourse, providing the baseline methodology against which low-resource European Portuguese council records can be structured.
- **ELITR Minuting Corpus** [63]: Curated explicitly to tackle the transition from raw transcripts to structured administrative minutes, this corpus contains 179 project meetings (approx. 120 hours / 1.1M tokens) recorded across English and Czech. Benchmarked using joint Longformer [11] and BART [50] setups, the dataset established a baseline for "action-item" and agenda-point boundary recovery, though models exhibited high insertion penalty biases when trying to group informal technical chatter [63].

Secondary and Conversational References

- **DialSeg_711** [88]: Derived from 711 multi-turn English customer service logs and open-domain chat networks, this dataset is utilized to test model stability against high conversational noise. Introduced alongside a BERT-based next-utterance prediction baseline, it achieved a benchmark P_k of 0.174 [88].
- **Molweni** [52]: A large-scale corpus of 10,000 multi-party dialogues sourced from the Ubuntu Chat Corpus. It is methodologically unique for pairing topic boundary labels with

explicit Segmented Discourse Representation Theory (SDRT) rhetorical dependency trees; baseline evaluations using Discourse-Aware Graph Networks demonstrated that explicit rhetorical edges improve boundary detection P_k by 3.2% over flat textual encoders [52].

- **STAC** [4]: Sourced from 45 transcribed chat logs of an online version of the board game *The Settlers of Catan*, this corpus is used in specialized literature to study deceptive, highly strategic, and non-cooperative topical shifts.
- **TIAGE** [85]: Comprising 4,100 goal-oriented human-to-human chats, TIAGE provides a niche benchmark for models that must simultaneously track topic shifts and distinct task-completion states.

2.5.3 Domain-specific Corpora

Beyond broad encyclopedic knowledge and unscripted dialogue, a third class of corpora focuses on highly structured, rigid domains where thematic boundaries are dictated by strict genre conventions:

- **SegNews** [54]: Comprising 27k news articles paired with section-specific heading summaries, this dataset was built to test end-to-end generative segmentation. Evaluated via the joint **SegNews** framework (a modified Transformer), it established a combined segmentation-summarization baseline ROUGE-L of 34.2 [54].
- **Terms-of-Service** [5]: A large-scale corpus of 74,000 online legal agreements. Characterized by high-entropy procedural boilerplate and strict structural nesting, it offers a sharp counterpoint to expository text. Evaluated using a hierarchical BERT classifier, the baseline established a macro-F1 of 0.771 for detecting topical paragraph transitions [5].
- **Manifesto Project** [60]: A multilingual collection of over 4,000 annotated political party manifestos. In this corpus, segmentation is operationalized as the isolation of "quasi-sentences" to quantify specific policy pledges; fine-tuned RoBERTa baselines achieved a thematic transition accuracy of 81.6% [60].

2.6 Critical Analysis and Research Gap

Despite the significant advancements in Text Segmentation driven by deep learning and LLMs, a critical analysis of the literature reveals several persistent gaps, particularly when transitioning from general-purpose English datasets to specialized, non-English administrative contexts. The specific challenges posed by European Portuguese city council minutes remain largely unaddressed, falling into three primary categories: the language gap, domain specificity, and the unique structural format of minutes.

Table 2.3: Comparative Overview of Primary Text Segmentation Datasets and Benchmarks

Dataset	Category	Language	Source Media	Scale	Target Granularity	Baseline	Year
WIKI-727K	General	EN	Wikipedia	727,746 docs	Hierarchical (TOC)	TextSeg (Bi-LSTM)	2018
WikiSection	General	EN, DE	Wiki (Cities/Disease)	38k docs	Linear (Section)	SECTOR	2019
ICSI Corpus	Spoken	EN	Academic Meetings	75 sessions (72 hrs)	Linear / Conversational	LCSeg	2004
AMI Corpus	Spoken	EN	Scenario Meetings	100 hours	Linear / Goal-Oriented	DeepTiling	2005
MeetingBank	Spoken	EN	Municipal Councils	1,366 sessions (3.5k hrs)	Agenda-Item (Minutes)	LLM Prompting / BART	2023
ELITR Corpus	Spoken	EN, CS	Project Meetings	179 sessions (120 hrs)	Action-Item / Minutes	Longformer+BART	2022
DialSeg_711	Dialogue	EN	Customer Support	711 dialogues	Linear Utterance	BERT-NUP	2021
Molweni	Dialogue	EN	Ubuntu Tech Chat	10,000 dialogues	Rhetorical (SDRT)	DAGN	2020
STAC	Dialogue	EN	Game Chat (<i>Catan</i>)	45 dialogues	Strategic Discourse	SDRT Parsers	2016
TIAGE	Dialogue	EN	Task Dialogues	4,100 dialogues	Task-State Shift	BERT Classifier	2021
SegNews	Domain	EN	News Publications	27k articles	Joint Seg-Summary	SegNews (Transformer)	2022
Terms-of-Service	Domain	EN	Legal Contracts	74k agreements	Functional Paragraph	Hierarchical BERT	2021
Manifesto Project	Domain	Multi.	Political Platforms	>4,000 documents	Quasi-sentence (Policy)	Fine-tuned RoBERTa	2021

2.6.1 The Language and Resource Gap

The vast majority of supervised Text Segmentation models are optimized for English, leveraging massive benchmarks like Wiki-727K [47] or WikiSection [3]. While multilingual encoders like mBERT or XLM-R offer a bridge, the landscape for European Portuguese remains remarkably sparse. Unlike English and German, there is currently no standardized, publicly available dataset for European Portuguese topic segmentation. As evidenced by the study of Freisinger et al. [27], research in in European Portuguese, typically conducted on private or ad-hoc corpora such as lecture transcripts, suffers from limited reproducibility and benchmarking potential.

Furthermore, their findings demonstrate that zero-shot transfers from English-centric models often fail to capture the specific syntactic, rhythmic, and prosodic cues of the Portuguese language. This underscores that without native datasets or domain-specific fine-tuning, current models cannot achieve the level of precision required for the rigorous structural demands of official administrative archives. Consequently, researchers in this space are forced to either rely on noisy cross-lingual transfers or undertake the resource-intensive task of manual annotation to capture the unique nuances of Portuguese municipal governance.

2.6.2 Administrative Domain Specificity

Although most contemporary models are benchmarked on Wikipedia, which is officially guided by the editorial principle of a neutral point of view (NPOV), adherence to this standard is highly variable in practice, resulting in a stylistically heterogeneous body of expository prose. In sharp contrast, administrative transcripts from city councils are defined by highly specific procedural language, such as formal voting protocols, agenda items and legalistic interjections, that differ fundamentally from encyclopedic text. General-purpose architectures, such as Cross-Segment BERT [56], are designed to detect broad thematic shifts but may fail to recognize the subtle "micro-shifts" in administrative procedure. The work of Aumiller et al. [5] in the legal domain demonstrates that when a model does not understand the functional role of a paragraph (e.g., distinguishing a "motion" from a "discussion"), its ability to accurately place boundaries is severely diminished.

Consequently, there is a clear need for models that can integrate domain-specific vocabulary and procedural knowledge into the segmentation process.

2.6.3 Challenges in Structuring Municipal Minutes

City council minutes represent a unique form of administrative writing that remains relatively underexplored in both linguistic and computational research. As a genre, they fall somewhere between the loosely structured, spontaneous nature of everyday conversation and the highly conventionalized, rigid structure of legal documents. This hybrid nature presents a significant obstacle for transfer learning. Ghosh et al. [34] provided a compelling critique of this issue, demonstrating that strategies involving pre-training on large structured corpora (like Wikipedia) do not effectively transfer to unstructured conversational data.

Moreover, since administrative minutes, particularly those derived from municipal council sessions, frequently serve as a record of oral proceedings, they often inherit the linguistic complexities of spoken discourse, such as disfluencies and non-linear thematic progression. This makes them susceptible to the "label imbalance" and "near-miss" problems discussed in the metrics section. There is a significant research gap in the lack of standardized, hierarchical segmentation models that can handle the specific semi-structured flow of a political meeting, where participants may revisit previous topics or digress into informal banter. Addressing this requires moving away from standard sequence labelling towards models that can handle the inherent noise of administrative discourse while respecting its formal procedural requirements.

2.7 Summary

In summary, this chapter established the theoretical and methodological foundations of Text Segmentation. By tracing the evolution of the field from foundational unsupervised models based on lexical cohesion to modern supervised neural architectures, it becomes evident that while sequence-labeling models achieve remarkable precision on structured, encyclopedic text, they face significant hurdles in noisy, real-world domains. The review highlighted that conversational and administrative transcripts introduce unique challenges, such as domain-specific jargon, disfluencies, and highly localized procedural formatting, that standard models struggle to resolve.

Crucially, this analysis also identified a significant research gap: the absence of domain-specific segmentation models and datasets for European Portuguese municipal discourse. Current architectures predominantly rely on English-centric pre-training and fail to seamlessly capture the hybrid procedural-deliberative nature of city council minutes. Having established this gap, the following chapter introduces the CitiLink-Minutes dataset, addressing the critical need for a standardized, privacy-compliant corpus to train and evaluate Text Segmentation models in this highly specialized municipal domain.

Chapter 3

The CitiLink-Minutes Dataset

This chapter introduces the CitiLink-Minutes dataset¹ [12], the primary corpus utilized throughout this research. The construction of this resource was a collaborative effort, culminating in a co-authored publication [12]. Built as part of the broader CitiLink project², it stands as the first multilayer-annotated corpus of city council minutes in European Portuguese. By providing a curated, privacy-compliant resource, this dataset facilitates the exploration of Natural Language Processing (NLP), Information Retrieval (IR), and civic technology applications within the municipal domain.

The remainder of this chapter is structured as follows: Section 3.1 details the challenges of municipal data, the data collection process, and the annotation framework. Section 3.2 outlines the programmatic pipeline used to construct the hierarchical resource. Finally, Section 3.3 characterizes the dataset’s structural statistics and specific relevance for text segmentation tasks.

3.1 Municipal Data Challenges, Collection, and Annotation

As identified in the preceding analysis of the state-of-the-art, research on institutional and governmental discourse has historically relied almost exclusively on parliamentary proceedings, which primarily capture spoken interaction. Official city council minutes represent a rich yet under-explored source of civic information. They document local decision-making processes, policy discussions, and voting outcomes, offering structured evidence of local government operations at the level where citizens most directly interact with democratic institutions.

However, progress in the computational analysis of these records has been hindered by a critical lack of curated, standardized, and openly available domain-specific corpora. The municipal domain poses unique challenges for NLP, which are conceptually summarized in Figure 3.1. First, heterogeneous formatting across municipalities means that documentation practices vary wildly; for instance, one municipality might record voting outcomes in strictly formatted tables, while another embeds the exact same information within dense, unstructured prose. Second, municipal

¹A public sample of 6 annotated documents is available in the official repository (<https://github.com/inesctec/citilink-dataset>).

²<https://projeto.citilink.inesctec.pt/>

minutes rely heavily on domain-specific terminology, utilizing localized administrative jargon and legal statutory references, such as *Agenda*, *Ratification*, or *Resolution*, that standard pre-trained language models struggle to parse accurately without context. Finally, these documents exhibit a complex textual interplay between procedural formality and substantive deliberation. A single text block frequently shifts abruptly between rigid legal boilerplate and the organic, sometimes informal summarization of a lively political debate.

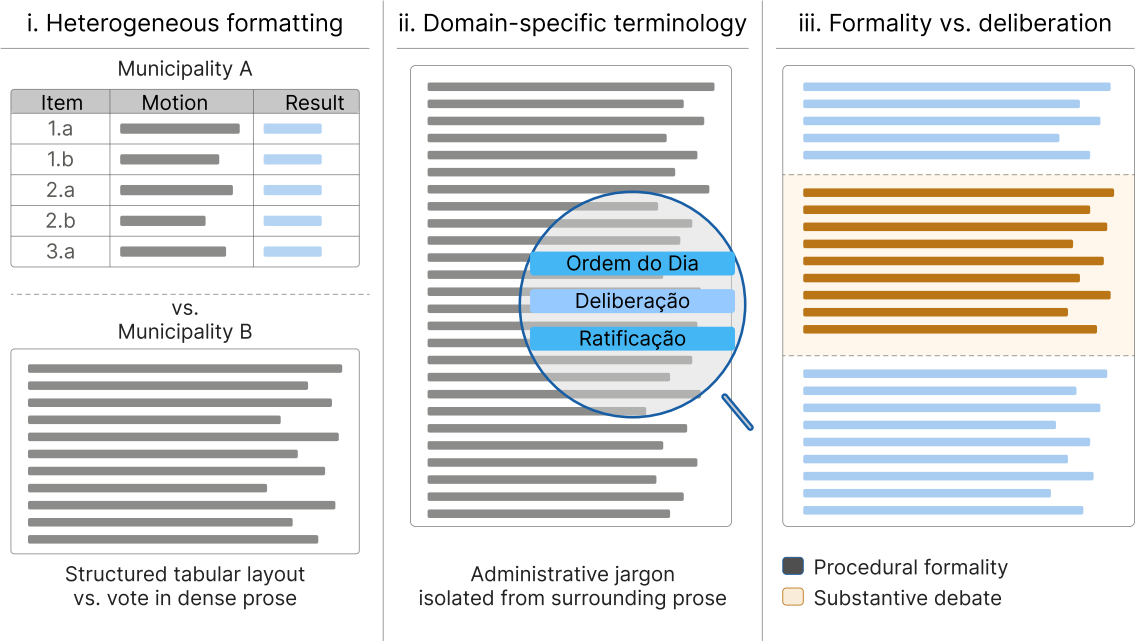


Figure 3.1: Primary NLP challenges in municipal documents: (i) heterogeneous formatting across different municipalities, (ii) the prevalence of domain-specific administrative jargon, and (iii) the abrupt textual shifts between rigid procedural formality and organic substantive deliberation.

To address these challenges, the CitiLink-Minutes dataset was developed. It comprises 120 official municipal minutes collected from six Portuguese municipalities: Alandroal, Campo Maior, Covilhã, Fundão, Guimarães, and Porto. These documents cover the 2021–2024 administrative term. The corpus was selected to ensure a representative diversity of session types (e.g., ordinary, extraordinary, public) while maintaining annotation feasibility. The dataset contains over 1.37 million tokens with comprehensive multilayer annotations covering (1) personal information, (2) metadata, (3) subjects of discussion, and (4) voting outcomes, totaling 44,859 entities and 11,366 relations. A detailed breakdown of these statistics per municipality is provided in Table 3.1.

The manual annotation process was conducted by a team of linguists using the INCEpTION platform [46]. The framework followed a modular architecture based on the ISO 24617 Semantic Annotation Framework (SemAF [42]), capturing information through complex structures of entities and links [12].

Before describing the annotation process, it is important to establish the foundational unit of analysis for the Text Segmentation (TS) task: the "segment," formally annotated as a Subject of Discussion (*Assunto*). A segment is defined as a distinct, self-contained deliberative topic.

Table 3.1: Corpus statistics for the CitiLink-Minutes dataset, distributed by municipality.

Municipality	Tokens	Entities	Relations
Alandroal	59,669	5,033	1,532
Campo Maior	155,436	7,130	1,905
Covilhã	298,496	11,922	2,809
Fundão	324,036	6,379	1,074
Guimarães	235,104	7,790	1,821
Porto	305,594	6,605	2,225
Total	1,378,335	44,859	11,366

These segments are hierarchically grouped under their respective parent Agenda Items (*Ordem do Dia*). Functionally, a segment encapsulates the entire lifecycle of a single discussion point, including the textual presentation, associated debate, and structured voting outcomes. By defining and extracting these segments, formal administrative documents are broken down into structured, searchable units that isolate individual subjects of discussion.

As illustrated in Figure 3.2, the annotations are distributed across four distinct conceptual layers. This multilayered architecture was specifically designed to organize the complex municipal data, thereby supporting downstream tasks such as information extraction, Personally Identifiable Information (PII) research, and Retrieval-Augmented Generation (RAG). The layers are structured as follows:

- **L1: Personal Information:** Captures instances that directly or indirectly identify individuals (which were subsequently anonymized to ensure privacy). These spans are explicitly categorized into distinct types such as PERSONAL-NAME, PERSONAL-ADDRESS, PERSONAL-ADMIN, PERSONAL-COMPANY, PERSONAL-POSITION, and PERSONAL-PUBLIC, directly supporting advanced PII detection.
- **L2: Metadata:** Records general session details, including participants, dates, locations, and meeting types (e.g., ordinary, extraordinary, public).
- **L3: Subjects of Discussion:** Defines the precise textual offsets (boundaries) for each subject of discussion. These subject boundaries constitute the primary ground truth labels for the Text Segmentation (TS) task developed in this dissertation. Each subject is semantically enriched with a free-text theme, categorized using a standardized 22-category municipal topic taxonomy (detailed in Table 3.3), and includes a summary field providing a concise natural language description of its core content. The above-mentioned team of linguists annotated the summaries, which were then extracted and integrated into the hierarchical JSON by the pipeline. This collaborative effort resulted in a publication introducing a new resource called **CitiLink-Summ** by Marques et al. [58], presented at The Web Conference (WWW) 2026. The contribution of this dissertation to that publication was limited to the foundational data engineering and pipeline extraction, whereas the summarization dataset

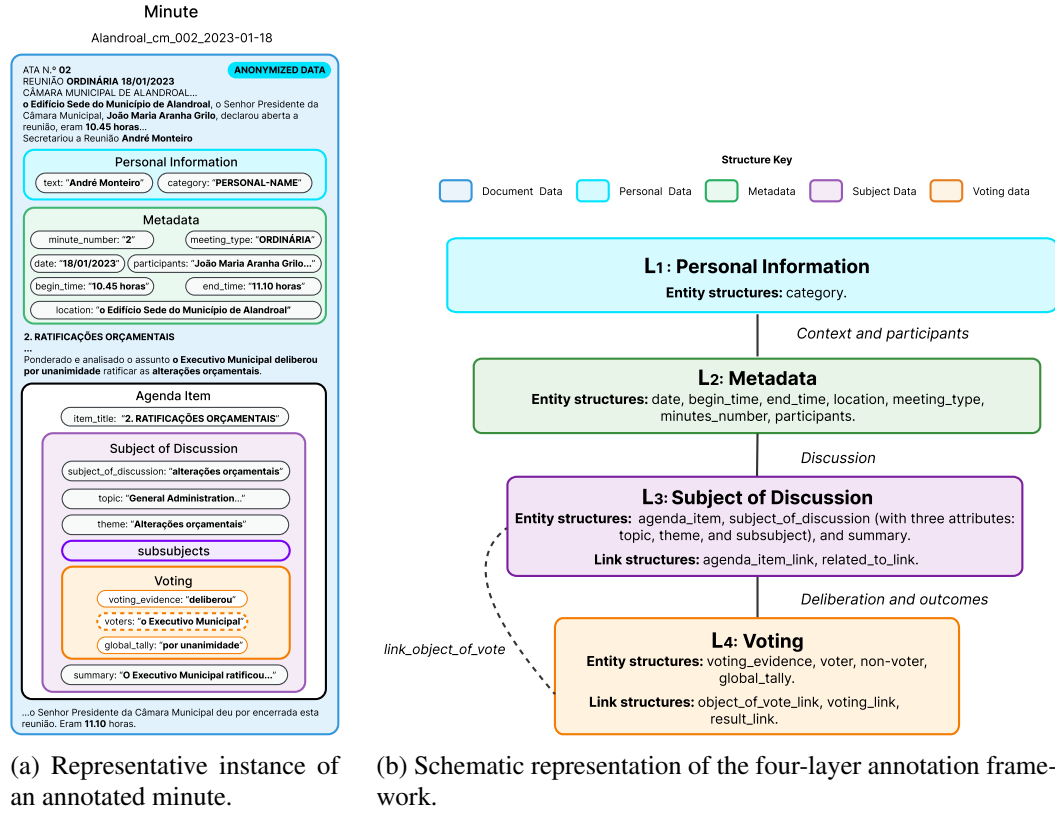


Figure 3.2: Overview of the four-layer annotation framework implemented in the CitiLink-Minutes dataset (adapted from Fig. 1 in [12]).

curation and analysis were led by the co-authors. Furthermore, to accurately reflect deeply nested municipal deliberations, a `sub_subsubjects` structure is also captured.

- **L4: Voting:** Captures the outcomes of votes, including voters, non-voters, and the global tally. This includes sub-voting records captured at the sub-subject level, exposing highly granular, per-item voting outcomes.

This chapter focuses primarily on the construction of the dataset. An exhaustive detailing of the semantic annotation scheme, linguistic guidelines, and inter-annotator agreement metrics is beyond the scope of this work. For a comprehensive description of the annotation methodology, readers are directed to the official resource paper [12] and the project’s publicly available Annotation Guidelines³.

3.2 Dataset Construction and Structural Transformation

While the linguistic team provided the initial annotations, translating these into a viable computational resource required a dedicated programmatic pipeline. The INCEPTION platform [46] ex-

³https://github.com/inesctec/citilink-dataset/blob/master/docs/citilink_annotation_guidelines.pdf

ports data in the Unstructured Information Management Architecture Common Analysis Structure (UIMA CAS) format. Under the ISO 24617 Semantic Annotation Framework (SemAF) specification utilized by INCEpTION, UIMA CAS represents annotations using stand-off markup (character offsets). As a result, the annotations are not natively hierarchical; instead, they exist as a flat graph of isolated entity nodes connected by relational links.

While UIMA CAS is highly effective for managing overlapping annotations during the manual labeling process, it is not easily interpreted by standard downstream NLP models. If the dataset were left in this exported state, researchers would have to independently engineer algorithms to resolve relational dependencies, such as determining which specific voting outcome belongs to which deliberative subject. Therefore, a conversion pipeline was required to traverse the relational graph, resolve the dependencies between nodes, and reconstruct the logical hierarchy of the municipal minute into a nested JSON structure.

This transformation process presented several challenges:

- **Nested and Overlapping Spans:** Municipal minutes frequently contain nested structures (e.g., a voting event containing references to multiple council members and a specific proposal), requiring careful handling of span boundaries during extraction.
- **Heterogeneous Relation Types:** The four annotation layers involve distinct relation semantics. For example, a single "Posicionamento" (Positioning) relation might link a specific voter to an "a favor" (in favor) stance, while a separate "resultado" (result) relation links the vote to a "por maioria" (by majority) outcome. These separate links must be programmatically matched and resolved into a unified voting object.
- **Graph Complexity:** The core challenge was building a system capable of recursively traversing the flat relational graph to find these specific semantic links, merge the connected entities, and rebuild the agenda-centric flow of the document.

To overcome these challenges and bridge the gap between raw linguistic annotation and downstream applications, a 6-step automated Python data pipeline⁴ was developed (Figure 3.3). The transformation process consists of the following sequential stages:

1. **Interval Extraction (Data Cleaning):** The pipeline first cross-references the raw annotations against predefined document boundaries. It removes invalid excerpts or non-deliberative content (marked by specific boundary annotations) and automatically adjusts all subsequent annotation offsets to maintain data integrity.
2. **Subject Parsing and Hierarchy Reconstruction:** This is the core transformation stage. The parser traverses the flat relational graph to extract subjects based on boundary markers, linking them to their respective agenda items ("Ordem do Dia"). It recursively maps voting methods, explicit voter positioning (in favor, against, abstention, blank), and global tallies, nesting them cleanly within individual subjects and sub-subjects. Additionally, it extracts

⁴<https://github.com/jmisidro/citilink-dataset-pipeline>

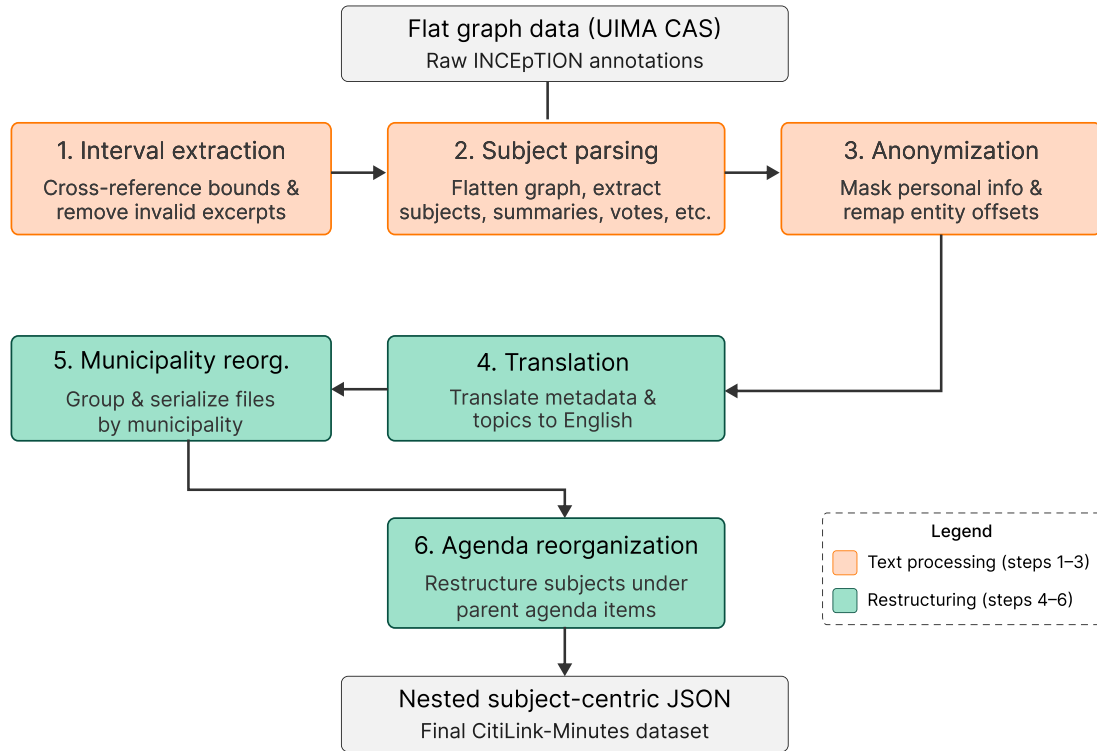


Figure 3.3: The 6-step automated data pipeline transforming raw INCEpTION annotations into the final subject-centric JSON hierarchy.

standardized topics, free-text themes (which act as descriptive titles), and natural language summaries for each subject. Finally, it also extracts personal information categorized into detailed types.

3. **Anonymization and Offset Remapping:** To ensure privacy compliance while maintaining the natural linguistic flow of the documents, categorized personal information is not merely masked but replaced with contextually accurate synthetic data. This process utilizes the Faker library [25] alongside custom generator functions to produce realistic Portuguese names, addresses, dates, vehicle descriptions, and administrative codes. The pipeline also employs a heuristic gender detection module to ensure that synthetic names and job titles grammatically align with the original text. Critically, replacing original text with synthetic strings of varying lengths (e.g., replacing a 4-character name with a 15-character synthetic name) inherently shifts the character positions of the entire document. If left unadjusted, all subsequent annotations would point to incorrect text spans. To solve this, the pipeline calculates the character length delta for every replacement and maintains a running cumulative offset. It then dynamically recalculates and remaps the `start` and `end` indices for all downstream entities, such as voting evidence, voter names, and global tallies, guaranteeing consistency across the final JSON dataset.

4. **Metadata Translation:** To broaden the dataset’s accessibility for the international research

community, domain-specific topics, free-text themes, and metadata fields are programmatically mapped and translated from Portuguese to English. This step is fully automated, with topic labels being translated via a predefined lookup table of the 22 standardized categories.

5. **Municipality Reorganization:** The processed documents are systematically grouped and serialized into isolated files corresponding to their originating municipality.
6. **Agenda Item Reorganization:** In the final step, the data is restructured into an agenda-centric hierarchy, grouping all deliberative subjects strictly under their parent agenda items to support targeted Information Retrieval tasks.

To illustrate the need of this pipeline, Listing 1 demonstrates a simplified excerpt of the exported UIMA CAS format for a document from Município de Guimarães. In this format, the voting evidence (ID 65, "Votante"), the overarching subject (ID 67), and the specific vote relations (IDs 1-9) exist as disconnected elements scattered across arrays. Listing 2 illustrates the conceptual output of the pipeline, where these relational dependencies have been programmatically resolved and grouped into a unified object.

```
{
  "%FEATURE_STRUCTURES": [
    {
      "%ID": 1, "%TYPE": "custom.Relation",
      "begin": 4533, "end": 4542,
      "@Dependent": 3, "@Governor": 4,
      "label": "posicionamento",
      "posicionamento": "a favor"
    },
    {
      "%ID": 65, "%TYPE": "custom.Span",
      "begin": 12069, "end": 12090,
      "label": "Posicionamento",
      "Posicionamento": "Votante"
    },
    {
      "%ID": 67, "%TYPE": "custom.Span",
      "begin": 12127, "end": 12203,
      "label": "Assunto",
      "Tema": "Legalizacao de projeto de arquitetura"
    }
  ]
}
```

Listing 1: Simplified excerpt of the raw INCEpTION flat graph output (UIMA CAS) for Município de Guimarães.

The culmination of this 6-step pipeline is the complete and standardized dataset. As depicted in Figure 3.4, the dataset is systematically organized into arrays and objects. Each document encapsulates the four annotation layers, where "Agenda Items" contain nested "Subjects of Discussion" (along with any "Sub-Subjects"), which in turn contain their respective summaries and "Voting" records.

```

{
  "subject_id": "Guimaraes_cm_019_2021-11-18_1",
  "text": "O Executivo Municipal aprovou, por unanimidade...",
  "theme": "Legalizacao de projeto de arquitetura",
  "voting": [
    {
      "voting_evidence": { "text": "Votante", "start": 12069, "end": 12090 },
      "voters": {
        "in_favor": [ { "text": "...", "start": 4533, "end": 4542 } ]
      },
      "global_tally": { "text": "por maioria", "type": "majority" }
    }
  ]
}

```

Listing 2: Simplified conceptual mock of the parsed subject-centric JSON output after Step 2.

This standardized JSON data structure (detailed in Listing 3) directly enables downstream models to efficiently parse and interpret administrative discourse, facilitating the development of civic technology applications⁵.

3.3 Relevance for Text Segmentation

The CitiLink-Minutes dataset provides an exceptionally strong alignment with the requirements of Text Segmentation (TS), precisely because the challenges that make municipal minutes difficult to process computationally are the same challenges that a segmentation model must learn to overcome. In municipal minutes, the flow of deliberation is dictated by the agenda (*Ordem do Dia*). However, the transitions between different agenda items and distinct subjects are often visually ambiguous, embedded within dense administrative prose, or inconsistently formatted across different municipalities, as illustrated in Figure 3.1.

The 6-step transformation pipeline described in Section 3.2 explicitly resolves this ambiguity by recovering the exact textual boundaries of `agenda_items` and `subjects`. By mapping these semantic boundaries to precise text spans, the dataset provides a high-quality, privacy-compliant ground truth for training and evaluating automated segmentation models without requiring any additional manual labelling effort.

Specifically, the **subject** annotations (L3) are essential for the standard Text Segmentation (TS) task, as they provide the exact boundaries required to identify thematic shifts and train models to partition continuous text into distinct deliberative topics. Furthermore, the **agenda item** annotations (often referred to as sections) are fundamental for the Hierarchical Text Segmentation (HTS) explored in Chapter 5. They provide the broader macro-structure necessary for models to learn and recover the nested, multi-level organization of municipal minutes. The remaining annotation layers, L1 (Personal Information), L2 (Metadata), and L4 (Voting), are not directly exploited by the segmentation models proposed in this dissertation, but they substantially enrich the dataset's value

⁵The full Guimarães JSON sample is available at: https://github.com/inesctec/citilink-dataset/blob/v1.1.0/sample_data/Guimaraes.json

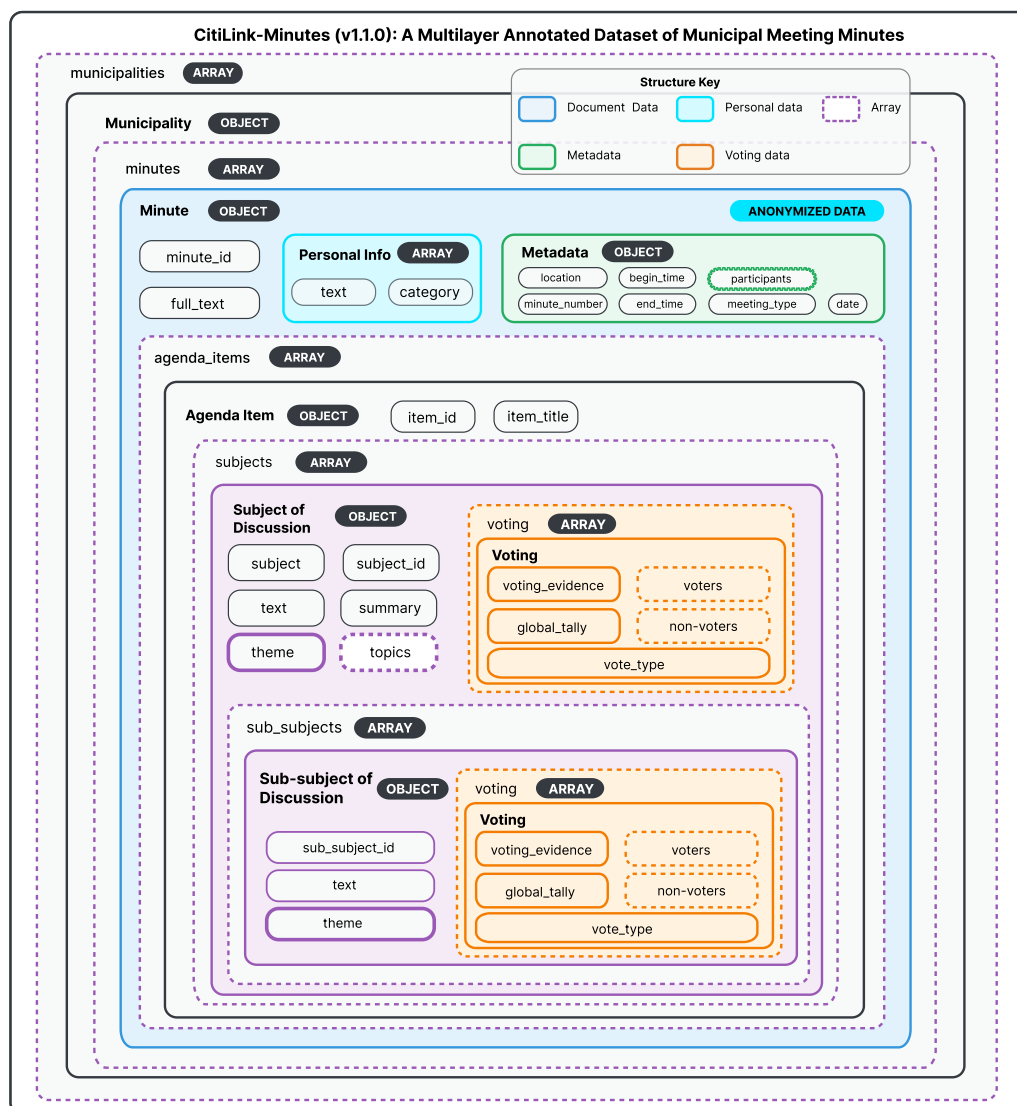


Figure 3.4: Summary of the JSON hierarchy for the CitiLink-Minutes dataset, showcasing the nested encapsulation of agenda items, subjects, sub-subjects, summaries, and voting data (adapted from Fig. 3 in [12]).

as a multi-purpose resource, enabling future work on **PII** detection, Named Entity Recognition, and civic **NLP** applications beyond the scope of this work.

To characterize the dataset in the context of this segmentation task, Table 3.2 presents the structural distribution and token statistics across the 120 minutes, detailed by municipality.

These statistics highlight both the structural granularity and the high variance of the dataset, both of which have direct implications for model design. For example, while Guimarães averages over 4,000 tokens per agenda item, Alandroal averages just over 100, which is a 40-fold difference that reflects the vast heterogeneity of municipal documentation practices. Similarly, average tokens per subject range from 72.98 (Alandroal) to 651.07 (Fundão), meaning that fixed-window segmentation approaches face a fundamental mismatch: a window calibrated for Fundão

Metric	Alandroal	Campo Maior	Covilhã	Fundão	Guimarães	Porto	Total
Total Municipal Minutes	20	20	20	20	20	20	120
Total Agenda Items	398	116	161	235	40	453	1403
Total Subjects	504	398	720	235 [†]	556	472	2885
Total Sub-Subjects	3	33	168	0	9	11	224
Average Subjects per Minute	25.20	19.90	36.00	11.75	27.80	23.60	24.04
Average Tokens per Subject	72.98	275.41	246.73	651.07	289.83	317.00	308.84
Average Tokens per Agenda Item	104.36	946.77	1113.34	668.00	4029.62	350.38	1202.08

Table 3.2: Corpus statistics characterizing the CitiLink-Minutes dataset, distributed by municipality; averages in the Total column represent the macro-average across the six municipalities. [†]Fundão’s annotation records one subject per agenda item, resulting in equal counts for Agenda Items and Subjects and no sub-subject nesting.

would span multiple complete subjects in Alandroal. This variance motivates the use of boundary-sensitive models capable of adapting to variable-length segments, a design consideration that informs the experimental choices described in subsequent chapters. Because the semantic boundaries are definitively anchored to the text spans regardless of these formatting and length discrepancies, computational models can be evaluated on their ability to accurately identify transitions between distinct deliberative subjects across diverse municipal styles. This establishes the dataset as the foundational resource for the Text Segmentation (TS) approaches proposed in this dissertation.

3.4 Summary

This chapter introduced the core annotation framework and detailed the structural transformation of the CitiLink-Minutes dataset. This is the first multilayer-annotated corpus of city council minutes in European Portuguese. Comprising 120 official minutes across six municipalities, the dataset exposes the vast heterogeneity and complex procedural language inherent to local governance. To bridge the gap between raw linguistic analysis and computational utility, a robust 6-step programmatic pipeline was developed to flatten dense relational annotations into a clean, hierarchical JSON schema.

Most importantly for this dissertation, this pipeline successfully recovered the precise textual boundaries of 1,403 agenda items and 2,885 nested subjects of discussion. By extracting both macro (agenda items) and micro (subjects) boundaries, the dataset establishes a reliable, privacy-compliant resource. This dual-level structural foundation directly enables the empirical evaluation of the Text Segmentation (TS) models developed in this research, setting the stage for the algorithmic frameworks and experimental results presented in the subsequent chapters. Specifically, the precise extraction of these subject boundaries serves as the fundamental ground truth required to train and evaluate SegNSP, the Linear Text Segmentation (LTS) framework introduced in the following chapter.

⁶https://github.com/inesctec/citilink-dataset/blob/master/docs/citilink_annotation_guidelines.pdf

Table 3.3: The 22-category municipal topic taxonomy used in the CitiLink-Minutes dataset to classify the semantic theme of subjects. For full definitions and examples, please refer to the official Annotation Guidelines⁶.

Topic	Description
General Administration, Finance, and Human Resources	Internal administrative management, financial and budgetary planning, human resources, public contracting, and anti-corruption measures.
Environment	Environmental management, sustainability policies, water and waste management, and preservation of natural resources.
Energy and Telecommunications	Promotion of energy efficiency, renewable energy adoption, and expansion of digital communication networks.
Traffic, Transport, and Communications	Planning and regulation of road traffic, public transit, mobility infrastructures, and road safety measures.
Education and Vocational Training	Management of school infrastructures, support for educational success, and promotion of vocational training programs.
Heritage	Preservation, management, and enhancement of the municipality's historical, cultural, tangible, and intangible heritage.
Culture	Promotion of cultural life, management of cultural facilities, and support for local artistic initiatives and events.
Science	Promotion of scientific knowledge, support for research projects, and partnerships with academic and scientific institutions.
Health	Public health policies, support for healthcare units, sanitary surveillance, and awareness programs.
Animal Protection	Animal welfare policies, management of shelters, responsible adoption promotion, and control of stray populations.
Sports	Promotion of physical activities, management of sports infrastructures, and support for local sports clubs and associations.
Social Action	Support for vulnerable populations, social solidarity initiatives, and the promotion of equality and inclusion.
Housing	Management of the municipal public housing stock, housing comfort programs, and allocation of municipal dwellings.
Civil Protection	Prevention of collective risks, disaster mitigation, and the operation of the Municipal Civil Protection Service.
Municipal Police	Creation and management of municipal police forces to maintain public order and enforce local regulations.
Public Works	Construction, safety, and maintenance of public infrastructures, roads, and municipal buildings.
Spatial Planning	Sustainable territorial development, land use planning, and Municipal Master Plan management.
Private Works	Management, licensing, and inspection of private property and urban development to ensure regulatory compliance.
Economic Activities	Local economic development, business establishment, local commerce stimulation, and agricultural initiatives.
External Cooperation and International Relations	Institutional relations, international networks, and cooperation agreements with intermunicipal or foreign entities.
Communication and Public Relations	Institutional communication, media relations, event organization, and administrative transparency initiatives.
Other	Matters not covered above, including public interventions, scheduling, and votes of praise or condolence.

```

{
  "municipalities": [
    {
      "municipality": "Guimaraes",
      "minutes": [
        {
          "minute_id": "Guimaraes_cm_019_2021-11-18",
          "metadata": {
            "date": "2021-11-18",
            "meeting_type": "Ordinaria",
            "participants": ["..."]
          },
          "personal_info": [
            {
              "text": "Tiago Correia",
              "type": "PERSONAL-NAME"
            },
            {
              "text": "Av de Ribeiro, Gandra",
              "type": "PERSONAL-ADDRESS"
            }
          ]
          // ... (remaining PII truncated)
        },
        "agenda_items": [
          {
            "subjects": [
              {
                "subject_id": "Guimaraes_cm_019_2021-11-18_1",
                "text": "...",
                "summary": "O executivo aprovou o novo regulamento de loteamento para a
                ↪ zona central...",
                "sub_subjects": [
                  {
                    "text": "...",
                    "voting": [
                      {
                        "vote_type": "Aprovacao",
                        "voters": ["..."],
                        "global_tally": "maioria"
                      }
                    ]
                  }
                ]
              }
            ]
          }
          // ... (remaining hierarchical data truncated)
        ]
      ]
    }
  ]
}

```

Listing 3: Abridged JSON schema sample (Guimarães) for CitiLink-Minutes v1.1.0

Chapter 4

SegNSP: A Linear Text Segmentation Framework

This chapter presents the proposed Linear Text Segmentation (**LTS**) solution, **SegNSP**¹, and details the main empirical findings. The content regarding this framework is adapted from our manuscript [43], currently under submission to EMNLP 2026. It outlines the architectural adaptations made to the Next Sentence Prediction (**NSP**) objective and evaluates the model’s capacity to navigate the structural variability of administrative texts.

Building upon the conceptual framing established in earlier chapters, we address the inherent ambiguity of topic transitions by focusing on sentence-level coherence, bypassing the architectural complexity of multi-task frameworks. To address these structural and efficiency gaps, this chapter introduces SegNSP [43], a streamlined, label-agnostic **NSP** framework.

Although **NSP** has largely been abandoned in modern language model pre-training (such as in RoBERTa) in favor of token-level objectives, its explicit modeling of sentence-to-sentence continuity makes it a natural candidate fit for detecting topic boundaries. By treating segmentation primarily as a sentence-level coherence task, SegNSP isolates the continuity signal between consecutive sentences. This provides an alternative that completely eliminates the computational overhead of auxiliary topic classification, allowing it to generalize to unlabelled administrative domains.

The remainder of this chapter is structured as follows. Section 4.1 details the problem formulation and the architectural adaptations made to the NSP objective. Section 4.2 outlines the experimental setup, including the datasets, baselines, and evaluation metrics. Section 4.3 presents a comprehensive evaluation of the model, including cross-domain performance, in-domain cross-municipality generalization, an ablation study on the loss function, an in-depth error analysis, and an assessment of computational efficiency and environmental impact. Furthermore, Section 4.4 validates the practical utility of the proposed segmentation framework through a downstream summarization task.

¹The source code, adapted dataset, and evaluation scripts for the SegNSP approach are publicly available at <https://github.com/jmisidro/segnsnp>.

4.1 Problem Formulation

Building upon this conceptual framing, a document D is represented as an ordered sequence of sentences (or utterances),

$$D = (s_1, s_2, \dots, s_n)$$

where each s_i denotes a single sentence. The objective of **LTS** consists of partitioning D into m contiguous, topically coherent segments

$$G = (seg_1, \dots, seg_m)$$

where each segment

$$seg_j = (s_{a_j}, \dots, s_{b_j})$$

is a subsequence of consecutive sentences discussing a common topic. Segment boundaries correspond to the transition points between adjacent segments, that is, between the last sentence of seg_j and the first sentence of seg_{j+1} .

Given this formulation, we cast linear text segmentation as a sentence-pair classification problem grounded in the **NSP** objective. For each pair of consecutive sentences (s_i, s_{i+1}) in a document D , a language model fine-tuned for NSP produces a probabilistic prediction

$$f_\theta : (s_i, s_{i+1}) \mapsto \hat{y}_i,$$

where $\hat{y}_i \in [0, 1]$ represents the likelihood that s_{i+1} continues the current topical segment. This framing models Text Segmentation (**TS**) as a boundary detection problem based on local discourse coherence, allowing the global segment structure to be reconstructed from sentence-level continuity signals.

We instantiate the function f_θ with a pretrained BERT model fine-tuned with the NSP objective. Each sentence pair (s_i, s_{i+1}) is encoded by concatenating the two sentences s_i and s_{i+1} and passing them through the model, producing a joint representation of the input sequence:

$$\mathbf{h}_{[\text{CLS}]} = \text{BERT}([\text{CLS}] s_i [\text{SEP}] s_{i+1} [\text{SEP}])$$

This representation is then fed into a linear classifier, which projects the high-dimensional sentence embeddings into a 2-dimensional space representing our target classes (`is_next` and `not_next`). The resulting logits are normalized with a softmax function to obtain NSP class probabilities:

$$P(y \mid s_i, s_{i+1}) = \text{softmax}(W\mathbf{h}_{[\text{CLS}]} + b),$$

where $y \in \{\text{is_next}, \text{not_next}\}$. During inference, a boundary is predicted at position i whenever the probability of the `not_next` class exceeds a decision threshold τ tuned on validation data, i.e.,

$$P(y = \text{not_next} \mid s_i, s_{i+1}) > \tau$$

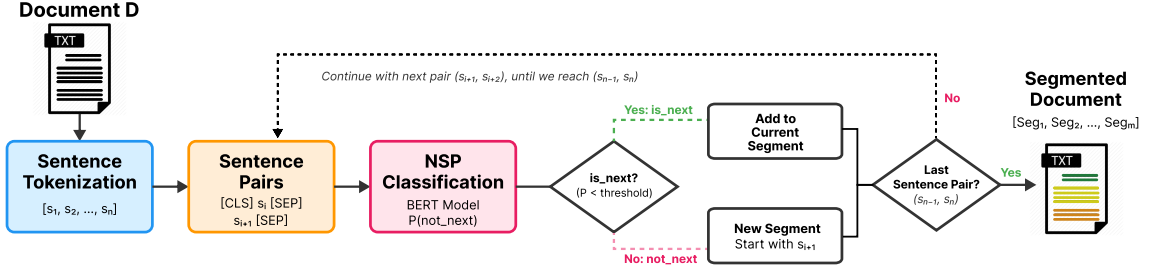


Figure 4.1: Overview of the NSP-based segmentation pipeline.

Contiguous spans between predicted boundaries then form coherent segments, reconstructing the document’s global topical structure. Figure 4.1 illustrates the end-to-end pipeline.

While the original BERT NSP formulation was designed for general pre-training and often relied on random sentence sampling, adapting it for linear text segmentation requires specific modifications to handle the sparsity and subtlety of topical boundaries. To this end, our approach introduces **domain-balanced ratios**, **harder negative sampling** and a **segmentation-aware loss**, effectively redefining NSP as a boundary-sensitive coherence classifier rather than a generic pre-training task:

- **Domain-Balanced Ratios:** In natural discourse, segment boundaries are sparse events. A model trained on raw consecutive pairs would mostly see "same-segment" examples, leading to conservative predictions and causing the model to frequently miss true topic transitions. By enforcing a fixed 70/30 distribution of intra-segment (positive) and inter-segment (boundary) pairs during batch generation, we artificially balance the learning signal. This ensures the model receives sufficient exposure to transition signals without being overwhelmed by the majority class, improving the learning of topical shifts.
- **Harder Negative Sampling:** Although the 30% inter-segment pairs provide the basic boundary signal, they often represent obvious shifts between distant topics. To improve robustness against "false boundary cues" and over-segmentation, we introduce "hard negatives" — up to ten randomly selected mismatched sentence pairs from within the same document (e.g., pairing a sentence from the beginning of a meeting with a non-adjacent sentence discussing a different procedural item later on). Unlike standard inter-segment pairs, which are adjacent, these "hard" negatives force the model to focus on local semantic continuity rather than superficial document-level cues, a strategy demonstrated to significantly improve discourse-level representations [44].
- **Segmentation-Aware Loss.** To further emphasise difficult boundary decisions, we employ a custom segmentation-aware loss function that combines three complementary objectives:

$$L_{\text{seg}} = w(\mathbf{si}, \mathbf{si} + 1) \cdot L_{\text{focal}} + \lambda \cdot L_{\text{conf}}, \quad (4.1)$$

where λ is a scalar confidence-penalty weight (default $\lambda=0.1$), and $w(\mathbf{si}, \mathbf{si} + 1)$ is a boundary-proximity salience weight described below. All three components operate on the same two-class logit vector $\mathbf{z} \in R^2$ produced by the BERT NSP head, with class 0 corresponding to *is-next* and class 1 to *not-next*.

Focal loss (L_{focal}). In the sentence-pair training data, intra-segment *is-next* pairs greatly outnumber inter-segment *not-next* pairs, creating a pronounced class imbalance. To address this, we replace standard cross-entropy with the focal loss [53]. For a sample with ground-truth label y and logits $\mathbf{z}$, let $\text{CE}(\mathbf{z}, y)$ denote the cross-entropy loss and define $p_t = \exp(-\text{CE}(\mathbf{z}, y))$ as the model’s probability of the correct class. The per-sample focal loss is:

$$\ell_{\text{focal}} = \alpha, (1 - p_t)^\gamma; \text{CE}(\mathbf{z}, y), \quad (4.2)$$

with $\alpha=1.0$ and $\gamma=2.0$ by default. The modulating factor $(1 - p_t)^\gamma$ continuously down-weights easy, well-classified examples (high p_t) while preserving the full gradient signal for hard cases (low p_t), precisely the ambiguous sentence pairs that cluster near topic transitions. L_{focal} is the mean of ℓ_{focal} over the mini-batch.

Confidence-penalty loss (L_{conf}). Administrative and council-meeting text contains many topic-ambiguous sentences that risk driving the model towards overconfident predictions on one class. To regularise this behaviour [70], we add a penalty that is proportional to the predicted confidence on misclassified samples:

$$L_{\text{conf}} = \frac{1}{B} \sum_{i=1}^B \hat{p}_{\max}^{(i)} \cdot \mathbf{1}[\hat{y}^{(i)} \neq y^{(i)}], \quad (4.3)$$

where B is the batch size, $\hat{p}_{\max}^{(i)} = \max_c; \text{softmax}(\mathbf{z}^{(i)})_c$ is the maximum softmax probability for sample i , and $\mathbf{1}[\hat{y}^{(i)} \neq y^{(i)}]$ is the binary indicator of a mis-prediction. A sample that is simultaneously wrong *and* highly confident incurs a large penalty, whereas a low-confidence mistake, or any correct prediction, contributes little or nothing to this term. This prevents the model from collapsing to trivial, overconfident predictions on the majority class.

Boundary-proximity weighting (w). Rather than a separate additive loss term, the boundary signal is incorporated multiplicatively through a per-sample salience weight w that scales L_{focal} in Equation (4.1). Each consecutive sentence pair $(\mathbf{si}, \mathbf{si} + 1)$ is assigned a weight during dataset construction, according to its proximity to the nearest ground-truth segment boundary:

$$w(s_i, s_{i+1}) = \begin{cases} 2.0 & \text{if the pair crosses a segment boundary} \\ 1.5 & \text{if the closest boundary is } < 3 \text{ sentences away} \\ 1.2 & \text{if the closest boundary is } < 5 \text{ sentences away} \\ 1.0 & \text{otherwise} \end{cases} \quad (4.4)$$

At the batch level the boundary contribution enters as $\bar{w} \cdot L_{\text{focal}}$, where $\bar{w} = \frac{1}{B} \sum_{i=1}^B w_i$ is the mean weight over the batch. This formulation assigns the greatest training emphasis to pairs that straddle a true boundary, moderate emphasis to pairs in the immediate vicinity, and baseline emphasis to pairs entirely within a segment, directly incentivising the model to prioritise the classification of challenging transition points. This is particularly important for boundary-based evaluation metrics such as the Boundary Edit Distance F-measure (B-F1) used for early stopping and final evaluation.

We empirically validate the necessity of each of these components through an ablation study in Section 4.3.3.

During training, the model learned to discriminate between sentence pairs that continue the current topical segment and those that span a boundary, under the adapted NSP formulation and segmentation-aware loss. At inference time, boundary probabilities $\hat{y}_i$ are computed for each consecutive sentence pair (s_i, s_{i+1}) and compared against a decision threshold τ to recover the document’s segmentation.

4.2 Experimental Setup

This section details the experimental framework developed to evaluate the SegNSP framework against alternative segmentation paradigms in administrative contexts. The methodology followed a rigorous protocol designed to ensure transparency, replicability, and the validity of the findings. In terms of the research questions, the methodology explicitly evaluated cross-domain and cross-lingual generalization (RQ1), and isolated the contribution of the NSP objective and its segmentation-aware enhancements (RQ2).

Datasets We evaluated our approach on two datasets that represent complementary evaluation scenarios. WikiSection [3] is a widely used benchmark for topic segmentation, derived from Wikipedia articles with explicit section boundaries. These boundaries provide a reliable approximation of topic shifts, making the dataset particularly suitable for segmentation evaluation. WikiSection exhibits relatively regular discourse, where topics are developed coherently and hierarchically. Its broad topical coverage (two categories: disease and city) and extensive use in prior work enable direct comparison with existing approaches, supporting reproducibility and fair benchmarking. In total, the dataset comprises approximately 38,000 full-text documents, over 242,000 annotated sections (segments), and 1.9 million sentences. In our experiments, we utilize all available subsets to explicitly test the cross-lingual capabilities and cross-domain generalization of our

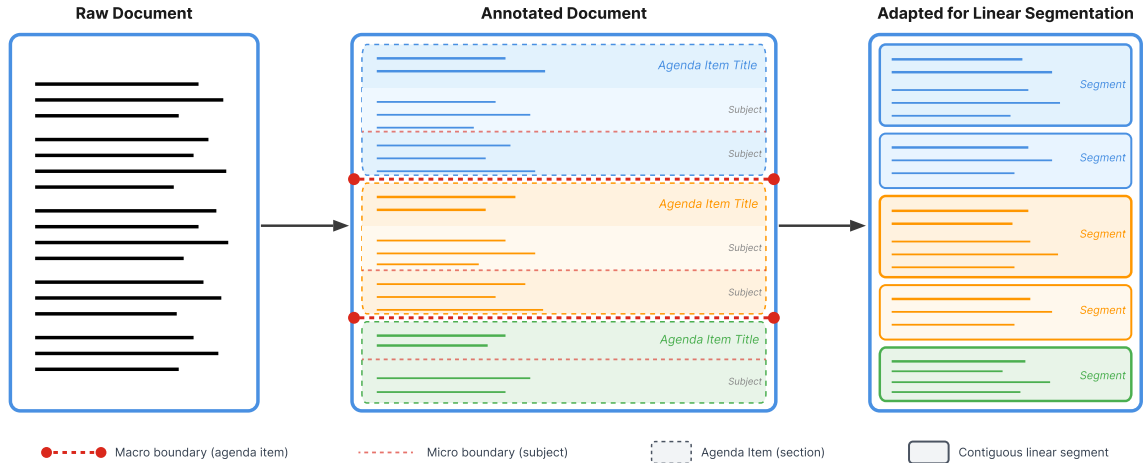


Figure 4.2: Adaptation of CitiLink-Minutes for linear text segmentation. Macro boundaries (agenda items) and micro boundaries (subjects) are merged to construct contiguous, non-overlapping segments.

approach: `en_city` (19,539 documents), `en_disease` (3,590 documents), `de_city` (12,537 documents), and `de_disease` (2,323 documents).

In contrast, CitiLink-Minutes [12] comprises 120 Portuguese city council meeting minutes collected from six municipalities between 2021 and 2024. Unlike WikiSection, topic boundaries in this corpus are often implicit, embedded in procedural language, enumerations, and shifts between administrative and deliberative content. Each municipality contributes 20 minutes, manually annotated with 2,885 subjects of discussion. To adapt this dataset for *LTS*, we flattened the structure by reconciling macro and micro boundaries. As illustrated in Figure 4.2, each agenda heading (macro boundary) was merged with the text of its first associated subject (micro boundary) to form a single initial segment. Subsequent subjects discussed under the same agenda item were maintained as distinct, standalone segments. This structural adaptation preserved the deliberative context while resulting in exactly 2,885 contiguous linear segments that map directly to the annotated subjects.

Together, these datasets provided a complementary evaluation framework. WikiSection enabled controlled benchmarking against existing methods, while CitiLink-Minutes tested model performance on realistic, less canonical discourse written in European Portuguese. Evaluating across both allowed us to assess the robustness and generalization of segmentation models across different domains, languages, and discourse structures.

Baselines We compared our NSP-based approach against several representative baselines. Text-Tiling [40] is a foundational unsupervised method that segments text by detecting lexical cohesion shifts between adjacent blocks, and has long served as a reference for discourse-level segmentation. Att+CNN [6] is a supervised neural model that combines convolutional and bidirectional LSTM layers with attention to predict sentence-level segment boundaries. TopSeg [32] is a coherence-driven model trained with contrastive learning, optimizing sentence embeddings so that

sentences within the same segment are closer in representation space, while sentences across segments are pushed apart. Topic boundaries are then identified where the cosine similarity between adjacent sentences falls below a threshold.

We also explored an alternative LLM-based segmentation paradigm inspired by *LumberChunker* [22], using Gemini 2.5 Flash Lite [18] as the underlying generative model. This allowed us to evaluate our approach in a cross-paradigm setup and assess its robustness against generative LLM-based segmentation. Other generative LLM-based methods (e.g., Malkiel et al. [57]) were not included, as they typically generate free-form text rather than producing explicit boundary offsets, preventing evaluation with boundary-based metrics. For all included baselines, we used publicly available implementations and recommended hyperparameters whenever possible to ensure fair and reproducible comparisons.

Metrics We evaluated segmentation quality using several complementary metrics. P_k [10] and *WindowDiff* (WD) [71] quantify boundary errors within a sliding window. P_k measures incorrect same-segment assignments, whereas WD is less punitive for near-miss boundaries. To mitigate the harsh treatment of near-miss boundaries, we additionally reported the boundary F_β -measure ($B-F_1$) and Boundary Similarity (B) [26]. $B-F_1$ treats segmentation as sequence labeling, granting partial credit for approximate matches, while B compares boundary sets directly, rewarding close alignments that P_k and WD may overlook. Because topic transitions can be gradual, the evaluation also considered tolerance windows, treating predictions near the gold boundary as partial matches.

Training and Evaluation Protocol Performance was assessed using a two-part evaluation protocol designed to capture both broad robustness and domain-specific adaptability. First, to evaluate general effectiveness and cross-domain generalizability, we employed standard train/test splits across the CitiLink-Minutes and WikiSection corpora. Second, to specifically quantify the model’s ability to generalize across the distinct local conventions of different administrative bodies, we utilized Leave-One-Out Cross-Validation (LOOCV) on the CitiLink-Minutes dataset, configured such that an entire municipality was omitted per fold. In each LOOCV fold, five municipalities were used for training and the remaining one for evaluation, ensuring that the model was tested on an entirely unseen administrative domain. This comprehensive setup enabled the assessment of SegNSP’s overall boundary detection capabilities alongside its specific ability to navigate the stylistic, structural, and topical variations present in municipal records.

4.2.1 Training Setup

We conducted experiments on CitiLink-Minutes and WikiSection to evaluate both **in-domain** and **cross-domain** performance. For WikiSection, we used the predefined standard splits of all subsets to evaluate cross-domain generalizability using standard train/test evaluation. For CitiLink-Minutes, we followed the chronological split of the corpus, assigning 60%, 20%, and 20% of documents for training, validation, and testing, respectively, reserving the most recent meeting minutes for testing to simulate realistic deployment conditions.

To accommodate the multilingual nature of our evaluation, SegNSP is initialized from language-specific pre-trained BERT encoders. Specifically, we utilized `bert-base-portuguese-cased`² [79] for the CitiLink-Minutes dataset, `bert-base-uncased`³ [21] for the English WikiSection subsets, and `gbert-base`⁴ [14] for the German subsets. The models were fine-tuned for sentence continuity prediction using the proposed segmentation-aware objective described in Section 4.1. Hyperparameters, including learning rate, batch size, and loss weighting coefficients, were optimized on the validation split, resulting in a learning rate of 5×10^{-6} , batch size of 8, focal parameters $\gamma=1.5$ and $\alpha=0.8$, confidence penalty of 0.15, and a boundary weight of 0.2, with early stopping applied after 12 epochs. Boundary detection during inference uses a threshold $\tau=0.5$, selected based on validation B- F_1 performance.

Training followed the intra/inter-segment ratios and harder negative sampling strategy described in Section 4.1, ensuring the model effectively captured both semantic continuity within segments and transitions across topics. All baselines were trained or tuned using their recommended configurations and publicly available implementations to ensure fair and reproducible comparisons.

4.3 Experiments and Results

In this section, we present a series of experiments designed to evaluate the performance of our proposed segmentation approach, **SegNSP**. We first assessed cross-domain and cross-lingual performance using standard train/test splits across the CitiLink-Minutes and WikiSection corpora to establish overall effectiveness. Subsequently, we examined in-domain cross-municipality generalization using the previously detailed LOOCV configuration on the CitiLink-Minutes corpus. Finally, we provide an ablation study parsing our proposed loss components, followed by a detailed error analysis.

4.3.1 Cross-Domain and Cross-Lingual Results

We evaluated **cross-domain** performance by comparing our approach against established baselines on CitiLink-Minutes and across all four subsets of WikiSection, as consolidated in Table 4.1. For error-based metrics (P_k and WindowDiff), lower scores indicate better performance (fewer boundary placement errors), whereas for sequence-labeling metrics (B- F_1 and B), higher scores approaching 100 signify superior alignment with ground-truth boundaries.

Our analysis of these results reveals several insights regarding the efficacy of different segmentation paradigms. On the administrative CitiLink-Minutes dataset, SegNSP achieved the highest B- F_1 (79.22) and Boundary Similarity B (59.31), indicating strong alignment between predicted and annotated topic transitions. Complementary error-based metrics reinforce this finding; an P_k of 8.07 means the model incorrectly grouped sentences within a sliding window only 8% of the time.

²<https://huggingface.co/neuralmind/bert-base-portuguese-cased>

³<https://huggingface.co/google-bert/bert-base-uncased>

⁴<https://huggingface.co/deepset/gbert-base>

Table 4.1: Cross-domain and cross-lingual segmentation results across Portuguese and English datasets using standard train/test splits (German subsets are relegated to Appendix A for readability). Higher $B-F_1/B$ and lower P_k/WD indicate better performance. Values are macro-averaged; best in bold.

Model	CitiLink-Minutes				en_city				en_disease			
	$B-F_1 \uparrow$	$B \uparrow$	$P_k \downarrow$	$WD \downarrow$	$B-F_1 \uparrow$	$B \uparrow$	$P_k \downarrow$	$WD \downarrow$	$B-F_1 \uparrow$	$B \uparrow$	$P_k \downarrow$	$WD \downarrow$
TextTiling	15.10	8.03	39.12	44.37	9.19	5.87	44.42	46.79	14.32	8.90	37.87	37.63
Att+CNN	34.00	21.00	25.00	27.00	43.44	29.23	23.80	29.47	14.60	8.64	36.45	37.13
TopSeg	42.00	26.00	18.00	25.00	31.32	18.45	15.90	19.48	23.00	13.17	30.03	33.51
LumberChunker	10.05	6.43	34.67	35.12	42.15	29.28	31.84	35.30	–	–	–	–
STP + TC + NSP	–	–	–	–	–	–	4.60	5.20	–	–	13.70	14.70
SegNSP	79.22	59.31	8.07	10.17	65.54	47.77	14.20	18.89	64.38	44.29	18.03	20.30

In this highly variable domain, traditional unsupervised methods like TextTiling performed poorly ($P_k = 39.12$), likely due to their strict reliance on explicit lexical cohesion, which is significantly undermined by the administrative boilerplate present in meeting minutes.

Crucially, the results from the WikiSection subsets provide compelling empirical evidence for the metric biases discussed in Section 4.2. While the TopSeg baseline achieves seemingly competitive error-based scores on the de_city subset ($P_k=18.53$, $WD=22.42$), its actual sequence-level boundary matching is exceptionally poor ($B-F_1=27.90$). In contrast, SegNSP not only establishes lower overall error rates ($P_k=17.02$, $WD=21.43$), but more importantly, achieves a $B-F_1$ score more than double that of TopSeg (60.09). This vast discrepancy perfectly exemplifies how window-based metrics can project a false sense of accuracy while masking poor boundary localization, reinforcing the necessity of sequence-labeling metrics like $B-F_1$ for fair evaluation.

Furthermore, we observed a stark contrast when evaluating the generative Large Language Model (LLM) paradigm. The LLM-based LumberChunker performed substantially worse on the structurally implicit CitiLink-Minutes ($B-F_1=10.05$) and did not improve over the strongest encoder-based baselines on the highly structured WikiSection en_city subset ($B-F_1=42.15$). This highlights the practical challenges of generative segmentation approaches for tasks requiring precise boundary detection, as LLMs often struggle to map generated chunks back to exact sentence offsets in the source text without hallucinating boundaries.

Finally, while the multi-task model by Lee et al. [48] (STP + TC + NSP) reported the lowest error metrics on en_city and en_disease, it was excluded from our sequence-labeling comparison due to its reliance on auxiliary topic labels and a lack of publicly available code for replication. Crucially, our local replication of established baselines (such as Att+CNN and TopSeg) on the WikiSection evaluation datasets yielded baseline error scores that are fully consistent with the original metrics documented in their respective literature, confirming the operational accuracy of our local benchmarking environment. However, the strong, highly transferable performance of our label-agnostic architecture directly addresses RQ2. It demonstrates that an NSP objective, when enhanced with harder negative sampling and a segmentation-aware loss, effectively captures local discourse continuity without the need for auxiliary topic classification. Overall, these results support SegNSP as a lightweight, encoder-based alternative that balances high boundary accuracy

with computational efficiency across diverse domains and languages.

4.3.2 In-Domain Cross-Municipality Generalization

To further assess in-domain robustness, we evaluated SegNSP on the CitiLink-Minutes dataset using a Leave-One-Municipality-Out Cross-Validation (LOOCV) framework across six distinct municipalities (Alandroal, Campo Maior, Covilhã, Fundão, Guimarães, and Porto). In this setup, five municipalities were utilized for training and the remaining one for evaluation, simulating deployment on a completely unseen administrative domain. It is important to note that this specific geographic and administrative breakdown is unique to the CitiLink-Minutes dataset. The Wiki-Section corpus lacks comparable hierarchical metadata, so a similar granular LOOCV analysis cannot be performed for it.

Table 4.2 shows noticeable variation in performance across municipalities, revealing how localized formatting conventions impact model transferability. SegNSP achieved strong results on several target domains, notably reaching a $B-F_1$ of 77.68 for Campo Maior and 74.31 for Alandroal. In these cases, the model successfully generalized broad administrative patterns, such as standard transitions between meeting open formalities and legislative voting.

Conversely, performance dropped drastically for Fundão ($B-F_1 = 24.83$) and Porto ($B-F_1 = 34.56$). This stark divergence highlights the inherent challenges of generalizing to domains with highly distinct, localized procedural conventions and stylistic drift. For instance, as explored in our qualitative error analysis, municipalities like Porto utilize highly compressed structural outlines or dense bullet-point enumerations that drastically shorten individual segment lengths compared to the training baseline.

Interestingly, looking at the error-based family of metrics on Fundão, the model maintains a relatively stable P_k (15.08) and WD (21.40) despite the low $B-F_1$ score. This behavior points to a high rate of near-miss boundaries, where the model recognizes local shifts in topical focus but fails to land on the exact sentence offset required by sequence-labeling metrics. These findings directly address RQ1 by proving that while Transformer-based encoders capture broad administrative discourse, municipality-specific threshold calibration or few-shot domain adaptation remains necessary to mitigate the impact of local formatting conventions.

Table 4.2: Leave-one-municipality-out cross-validation results on CitiLink-Minutes. Higher $B-F_1$ and B, and lower P_k and WD indicate better performance. Best values are in bold.

Municipality	$B-F_1 \uparrow$	B $\uparrow$	$P_k \downarrow$	WD $\downarrow$
Alandroal	74.31	49.46	8.24	9.35
Campo Maior	77.68	55.97	9.14	13.25
Covilhã	72.43	50.70	10.09	12.43
Fundão	24.83	15.92	15.08	21.40
Guimarães	67.93	48.76	7.12	12.28
Porto	34.56	19.85	11.30	19.88

4.3.3 Ablation Study

To isolate the contributions of each component of our segmentation-aware loss objective, we conduct a controlled ablation study on the CitiLink-Minutes dataset, analysing the impact of progressively removing components of $L_{\text{seg}} = \bar{w} \cdot L_{\text{focal}} + \lambda L_{\text{conf}}$ (Equation 4.1). The results are consolidated in Table 4.3.

Table 4.3: Ablation of loss components on CitiLink-Minutes.

Variant	B-F ₁ ↑	B ↑	P _k ↓	WD ↓
Full model ($\bar{w} \cdot L_{\text{focal}} + \lambda L_{\text{conf}}$)	79.22	59.31	8.07	10.17
w/o L_{focal} & L_{conf} (CrossEntropyLoss)	76.45	55.78	12.10	15.24
w/o L_{conf} ($\bar{w} \cdot L_{\text{focal}}$ only)	74.40	53.91	15.85	18.95
w/o boundary weight ($L_{\text{focal}} + \lambda L_{\text{conf}}$, $w=1$)	74.15	53.67	15.43	18.67
L_{focal} only ($w=1$, $\lambda=0$)	73.02	52.82	16.56	19.78

The ablation trajectory demonstrates that the baseline cross-entropy framework is structurally insufficient on its own for precise linear text segmentation, and that the full segmentation-aware loss combination is critical to boundary accuracy. Relying on focal loss alone (all $w_i=1$, $\lambda=0$) yields the weakest B-F₁ of 73.02 and the highest window error rates ($P_k=16.56$, WD=19.78), underscoring that simply managing class imbalance is insufficient when the model faces complex, locally continuous discourse.

Switching to standard cross-entropy (which disables both the focal modulation and the confidence penalty, since both are encapsulated in `SegmentationAwareLoss`) causes a drop in B-F₁ from 79.22 to 76.45 and a sharp increase in window errors (P_k rising from 8.07 to 12.10). This confirms that without an explicit mechanism to penalise easy negatives, the model defaults to conservative, over-smoothed predictions that frequently miss rare boundary events.

More tellingly, removing either the confidence penalty alone (w/o L_{conf}) or the boundary-proximity weighting alone (w/o boundary weight) degrades performance identically, dropping the B-F₁ to approximately 74.15–74.40 and driving WindowDiff to nearly 19. These two components address orthogonal failure modes: the confidence penalty prevents the model from making overconfident incorrect assertions on ambiguous, stylistically noisy text, while the boundary-proximity weight acts as a spatial smoother, assigning greater training emphasis to sentence pairs that straddle or closely approach a true boundary (Equation 4.4). Collectively, these results demonstrate that all components synergise to transform a generic pre-training task into a highly precise, boundary-sensitive sequence classifier.

4.3.4 Error Analysis

To move beyond aggregate performance metrics, we conducted a qualitative error analysis across both datasets. While WikiSection served as a controlled, established benchmark, CitiLink-Minutes reflected the structural variability and administrative boilerplate typical of real-world local governance records. Importantly, the analyses presented in this section are computed on the standard chronological train/test split, where all six municipalities are present in the training data. They

therefore characterize in-distribution boundary error patterns and should not be conflated with the cross-municipality generalization results of the LOOCV experiment (Section 4.3.2).

We first examined the distribution of $P(\text{not_next})$ across all consecutive sentence pairs. On CitiLink-Minutes (Figure 4.3), the model exhibited a sharply bimodal distribution, with only 1.4% of predictions falling within the ambiguous interval $[0.4, 0.6]$. In contrast, WikiSection (Figure 4.4) showed a broader overlap (4.7%) and a negative separation gap (-0.015), reflecting greater uncertainty in boundary decisions. This pattern suggests that semantic transitions in Wikipedia articles are often more gradual and less structurally explicit than in administrative meeting minutes.

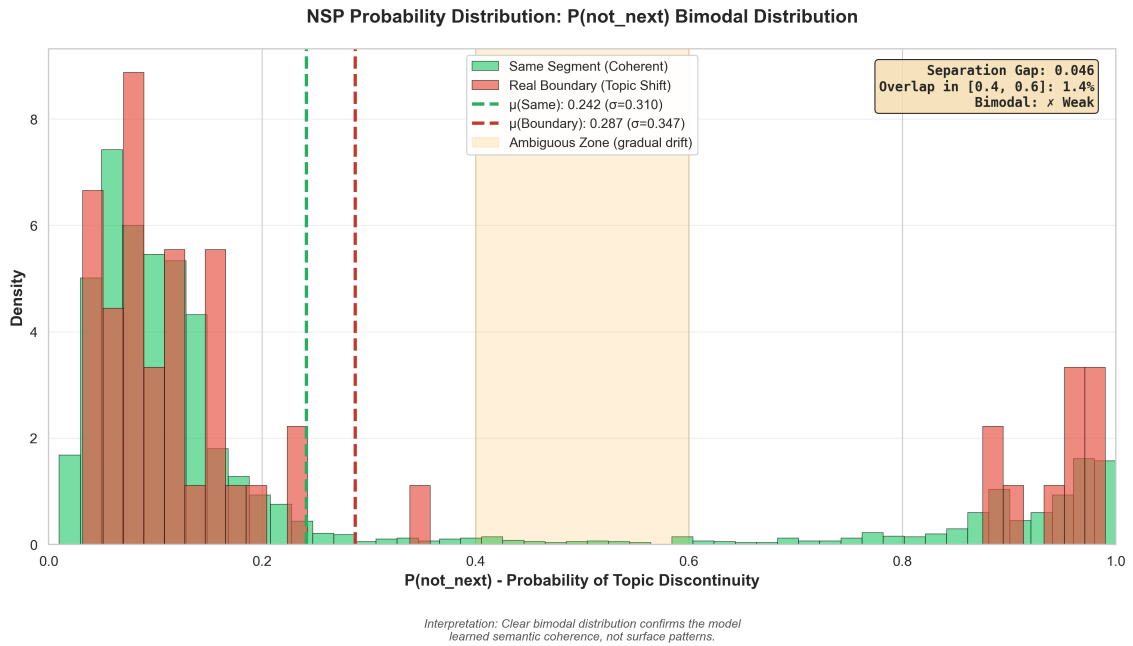


Figure 4.3: CitiLink-Minutes: bimodal distribution of $P(\text{not_next})$, showing a 1.4% overlap in the ambiguous zone $[0.4, 0.6]$.

A more granular analysis of positional error distributions revealed further dataset-specific trends. In CitiLink-Minutes (Figure 4.5), over-segmentation (false positives) increased toward the final 10% of documents, corresponding largely to administrative boilerplate, such as adjournment formalities, that is misclassified as topical transitions. Conversely, WikiSection exhibited a predominance of under-segmentation errors (false negatives) near document endings (Figure 4.6), consistent with shorter sections and more gradual topic transitions.

Examining errors at the municipality level (Figure 4.7) highlighted local stylistic effects within CitiLink-Minutes. Alandroal showed a higher rate of under-segmentation ($\text{FN} \approx 0.065$), indicating a conservative boundary detection that may miss subtle topic shifts. Campo Maior, by contrast, exhibited higher over-segmentation rates ($\text{FP} > 0.05$), potentially influenced by repetitive linguistic patterns. Fundão and Porto displayed near-zero error rates, indicating that the chosen decision threshold τ generalized effectively to these municipalities.

Overall, these analyses indicate that segmentation performance is shaped not only by domain

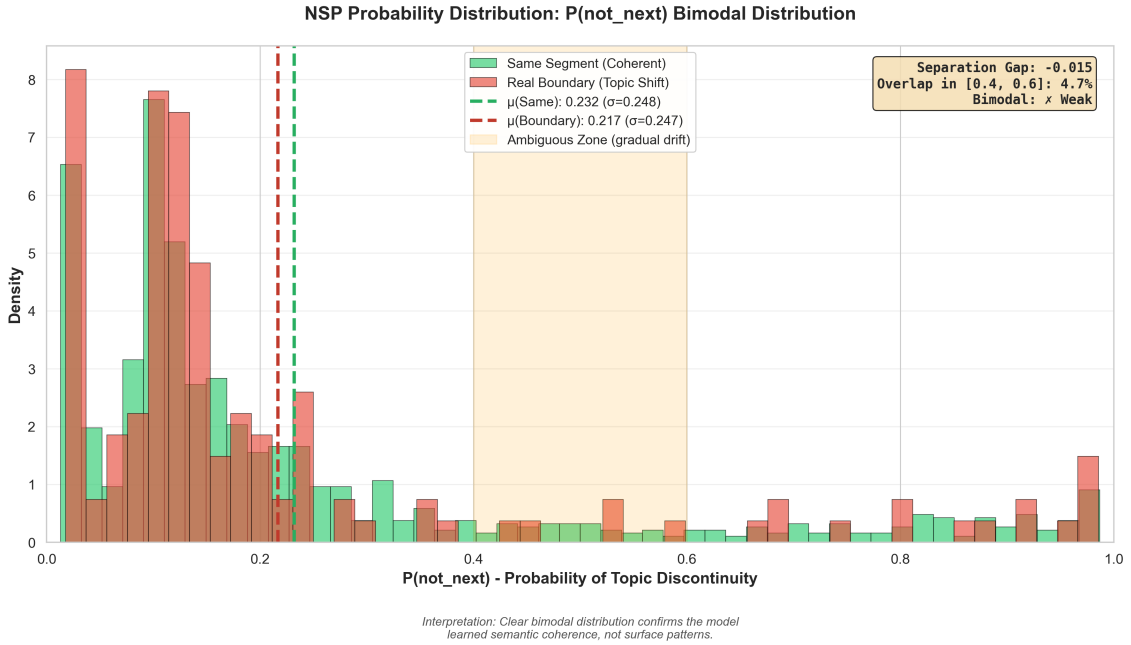


Figure 4.4: WikiSection: bimodal distribution of $P(\text{not_next})$, broader overlap (4.7%) in the ambiguous zone $[0.4, 0.6]$.

characteristics but also by localized stylistic conventions and structural variability. These findings motivate future work on adaptive thresholding or municipality-aware calibration to improve robustness across heterogeneous administrative records.

4.3.5 Computational Efficiency and Environmental Impact

A primary motivation for utilizing encoder-based models like SegNSP is their alignment with Green AI principles, offering significant reductions in carbon emissions and computational overhead relative to large language models. To substantiate this, we tracked our local footprint using the CodeCarbon library [20]. Executing our complete SegNSP pipeline on a local NVIDIA A100 GPU required approximately 8,224 seconds (2.28 hours), consuming 0.93 kWh of energy. Benefiting from a low-carbon energy grid, this generated an environmental footprint of just 0.00366 kg of CO₂e (3.66 grams).

To place this figure in perspective against established literature, standard fine-tuning benchmarks for BERT-class models tracked via CodeCarbon typically consume 3.1 kWh, emitting roughly 0.8 to 1.2 kg of CO₂e under standard grid intensities. Furthermore, parameter-efficient fine-tuning (PEFT) of a lightweight 7B parameter LLM on a highly renewable European grid produces approximately 0.057 kg of CO₂e. At just 3.66 grams of CO₂e, SegNSP’s total operational footprint is less than the emissions generated by charging a single modern smartphone (≈ 5 grams), positioning its environmental cost at a fraction of standard generative adaptation baselines.

In stark contrast, zero-shot generative segmentation using the Gemini 2.5 Flash Lite LLM

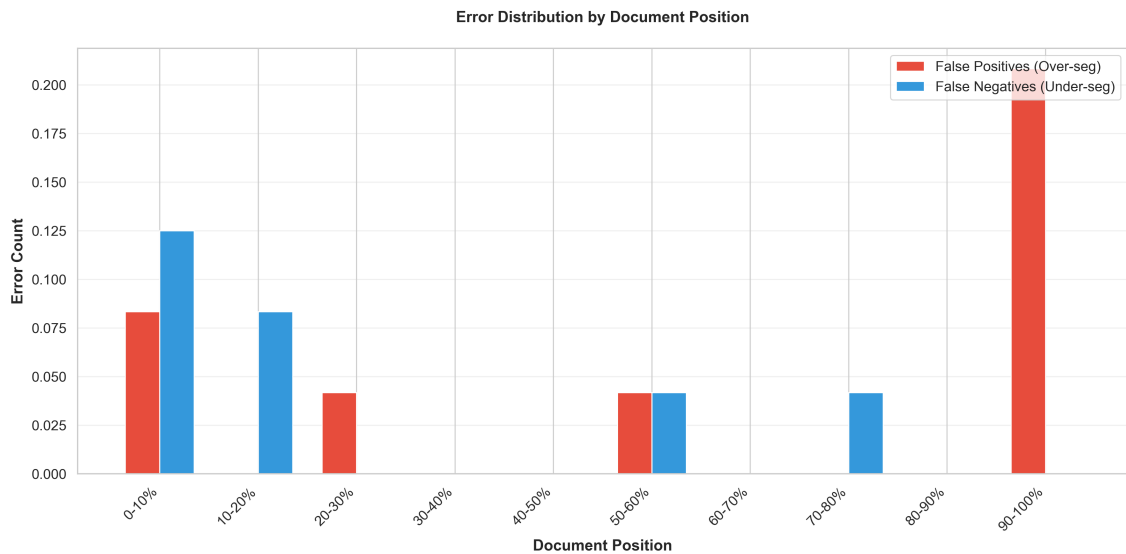


Figure 4.5: CitiLink-Minutes: error distribution across document positions.

(LumberChunker) inherently contradicts sustainable NLP practices for this specific task. LumberChunker operates sequentially, utilizing the predicted boundary of the current window to establish the start of the next. Segmenting our test dataset of 24 documents necessitates over 800 individual network API queries, processing an estimated 4 million tokens. While proprietary server-side emissions are opaque and cannot be directly measured by local tools like CodeCarbon, established literature on LLM inference demonstrates that routing thousands of sequential queries through a massive generative architecture produces a carbon footprint orders of magnitude larger than training and evaluating a 110M-parameter encoder.

Beyond the environmental cost, this reliance on cloud LLMs introduces severe latency bottlenecks; processing just the test set requires over 1.6 hours of continuous sequential inference due to API response times and rate limits, whereas SegNSP can infer boundaries for the same documents in seconds after training.

It is crucial to note, however, that this critique is strictly scoped to the task of **LTS**. While emerging local, open-weight **LLMs** (e.g., in the 8B to 12B parameter range) offer promising pathways to mitigate cloud reliance, data privacy concerns, and external API latency, the fundamental inefficiency of using auto-regressive generation for local sentence-to-sentence continuity remains. For this specific low-level classification task, deploying generative architectures, whether cloud-based or local, remains computationally disproportionate compared to lightweight encoders.

The computational calculus shifts entirely when addressing higher-order tasks. Recovering deeply nested, hierarchical document structures requires a level of global reasoning, semantic abstraction, and deductive logic that lightweight encoders struggle to achieve. In those complex scenarios, the advanced capabilities of LLMs present a compelling trade-off that warrants the thorough investigation undertaken in the subsequent chapter.

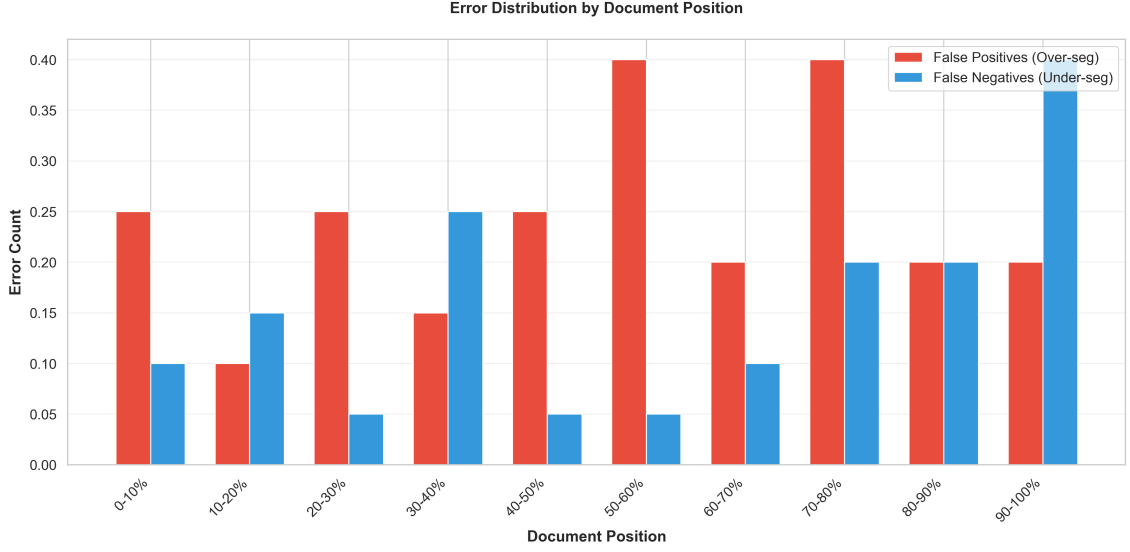


Figure 4.6: WikiSection: error distribution across document positions.

4.4 Downstream Task Evaluation: Summarization

While boundary detection metrics such as $B-F_1$ provide a robust mathematical evaluation of segmentation accuracy, text segmentation is rarely an end in itself. In modern Natural Language Processing (NLP) pipelines, segmentation serves as a structural prerequisite for downstream tasks. By dividing lengthy, dense documents into coherent semantic units, effective segmentation mitigates the "lost-in-the-middle" phenomenon and reduces hallucinations in LLMs processing long contexts. Therefore, to truly validate the utility of SegNSP, it is important to evaluate the quality of the downstream applications that rely on its predicted segments.

To demonstrate this, we conducted a downstream summarization experiment. We hypothesize that if SegNSP successfully captures the latent topical flow of the municipal minutes, the resulting predicted segments will contain sufficient semantic cohesion to support high-quality abstractive summarization, yielding results comparable to those generated from manually annotated boundaries.

4.4.1 Experimental Setup and Baseline Selection

For this evaluation, we replicated the summarization benchmark established by the **CitiLink-Summ** dataset [58], which introduced the first benchmark for summarizing discussion subjects in European Portuguese municipal minutes. While we co-authored this resource paper, our specific contribution was limited to the underlying dataset construction and data engineering; the summarization modeling and evaluation frameworks were developed entirely by the lead author. Therefore, we explicitly adopted the exact summarization pipeline and evaluation scripts sourced from the official CitiLink-Summ repository⁵, ensuring a strict, 1-to-1 methodological comparison

⁵<https://github.com/inesctec/citilink-summ>

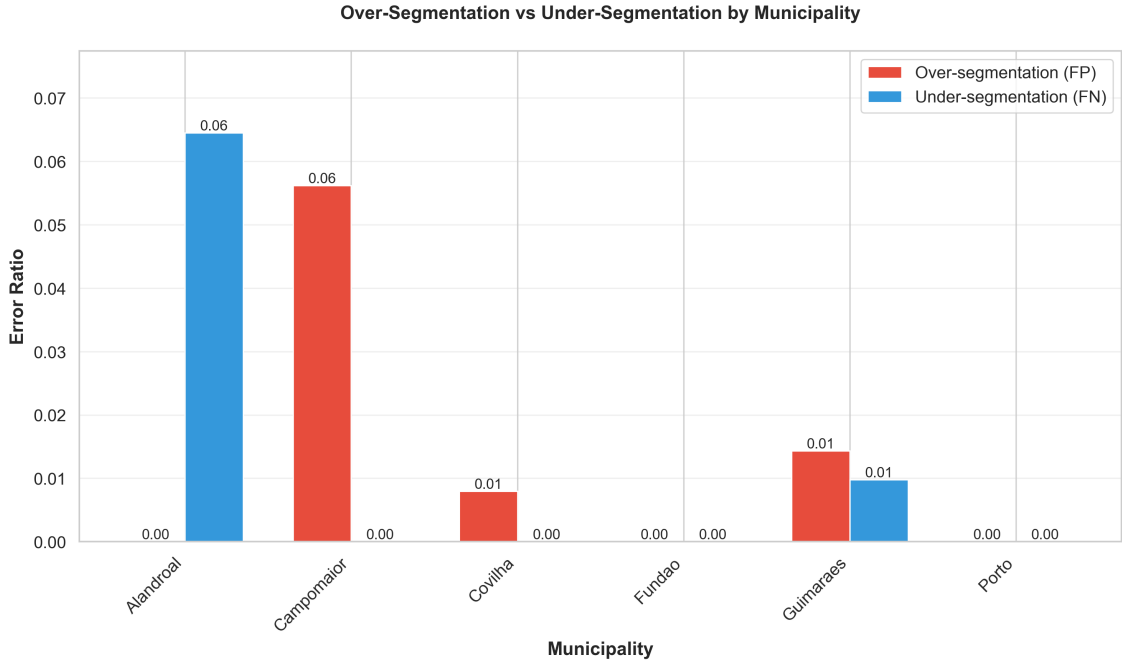


Figure 4.7: Over- and under-segmentation ratios across CitiLink-Minutes municipalities.

that isolates the impact of our segmentation model.

Following the CitiLink-Summ benchmark, we selected Gemini 2.5 Flash [18] as our generation baseline. The rationale for this choice is twofold. First, the original paper demonstrated that Gemini 2.5 Flash serves as a highly capable LLM baseline for this specific administrative domain. Second, because our objective is to evaluate the quality of the segmentation rather than the architecture of the summarizer, holding the generative model constant allows us to isolate the impact of the input text boundaries.

The experiment compares two distinct input scenarios:

1. **Reference Boundaries (Control):** Summaries generated by feeding Gemini 2.5 Flash the exact, manually annotated subject text spans from the dataset. (Results as reported in Marques et al. [58]).
2. **SegNSP Boundaries (Proposed):** Summaries generated by feeding Gemini 2.5 Flash the automated, predicted segments outputted by our SegNSP model.

The generated summaries were evaluated against the CitiLink-Summ manual reference summaries using the exact same suite of lexical and semantic metrics: ROUGE (1, 2, L), BLEU, METEOR, and BERTScore (Precision, Recall, F_1).

4.4.2 Results and Analysis

As presented in Table 4.4, the summaries generated from the SegNSP predicted boundaries exhibit lower scores across both lexical and semantic metrics when compared to the reference control

Table 4.4: Comparison of downstream summarization performance (scaled 0–100) using Gemini 2.5 Flash. The control group utilizes perfect manual annotations, while the proposed group relies entirely on SegNSP predicted boundaries.

Input Segment Source	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore (P)	BERTScore (R)	BERTScore (F ₁)
Reference Boundaries [58]	64.16	48.94	55.97	28.40	54.34	83.09	82.99	83.19
SegNSP Boundaries	35.99	17.28	28.72	9.34	29.49	69.05	73.41	70.95

group. This degradation is a natural consequence of error propagation; when a linear segmentation model misaligns a boundary (a "near-miss"), the generative **LLM** may occasionally process truncated context, absorb noise from an adjacent topic, or split a single cohesive deliberation across two separate summaries.

However, the critical takeaway is that despite lacking perfect boundary annotations, the semantic retention of the SegNSP results remains robust. A BERTScore F₁ of 70.95, equivalent to 85.3% of the reference ceiling of 83.19 ($70.95 / 83.19$), indicates that the semantic core of the generated text still closely aligns with the reference summaries, even if exact lexical matches (as measured by ROUGE and BLEU) suffer due to the shifted text windows. These findings provide strong empirical support for our hypothesis: SegNSP successfully encapsulates the necessary deliberative context, demonstrating that it is a viable preprocessing framework that actively supports downstream generative NLP tasks in unstructured, domain-specific discourse.

4.5 Summary

This chapter introduced SegNSP, a lightweight, label-agnostic framework that reframes Linear Text Segmentation (**LTS**) as a Next Sentence Prediction (**NSP**) task. To overcome the sparsity of topic transitions and the noise of administrative discourse, the standard **NSP** objective was significantly enhanced with domain-balanced training ratios, harder negative sampling, and a custom segmentation-aware loss function.

Through empirical evaluation, SegNSP demonstrated robust boundary detection capabilities. On the CitiLink-Minutes dataset, the model achieved a B-F₁ score of 79.22, a +37.22 point margin over TopSeg (B-F₁ = 42.00), the strongest alternative baseline evaluated, proving its capacity to navigate the structural variability of European Portuguese administrative texts. Furthermore, multilingual evaluations across the independent English and German WikiSection benchmarks confirmed that this architecture generalizes well across diverse canonical domains without the need for auxiliary topic classification.

Beyond boundary detection, the framework’s practical viability was validated through both computational and downstream evaluations. Profiling its environmental footprint demonstrated that the lightweight encoder architecture offers a highly sustainable, low-latency alternative to massive generative **LLMs** for local sentence-to-sentence continuity tasks. Moreover, a downstream summarization experiment confirmed that the segments predicted by SegNSP retain the semantic cohesion necessary to support coherent abstractive generation, yielding semantic retention

scores (BERTScore $F_1 = 70.95$) that maintain 85.3% of the upper-bound performance achieved using perfect manual reference boundaries.

Despite these strong results, the linear segmentation paradigm inherently assumes a flat document structure, treating all topical transitions with equal weight. As observed in the native structure of the CitiLink-Minutes dataset, real-world administrative discourse is nested, with specific subjects of discussion dynamically grouped under overarching agenda items. By flattening this hierarchy into a single linear sequence, models like SegNSP successfully identify immediate topical shifts but fail to capture the broader macro-structure that contextualizes related deliberations. This limitation highlights a critical gap in the field: the necessity of recovering the multi-level organization of complex documents. To address this structural shortcoming, the following chapter introduces a Hierarchical Text Segmentation (**HTS**) framework, moving beyond flat boundary detection to explicitly model and reconstruct the nested dependencies of municipal records.

Chapter 5

HSeg: A Two-Stage Hierarchical Text Segmentation Framework

This chapter introduces **HSeg**¹, a generative framework designed to capture the deeply nested discourse structure of administrative texts. Transitioning from the linear boundary detection explored in Chapter 4, it details a two-stage generative framework for Hierarchical Text Segmentation (**HTS**) and evaluates its performance across both domain-specific and open-domain benchmarks.

Historically, Text Segmentation (**TS**) has been framed as a flat, linear problem: locating transition points that divide a sequence of sentences into topically coherent spans [16, 40]. While Linear Text Segmentation (**LTS**) is well-studied, administrative meeting minutes and long-form encyclopedic articles routinely exhibit a rich, multi-level discourse structure. Broad procedural umbrellas, such as municipal agenda items, routinely contain multiple thematically distinct sub-deliberations. Forcing a flat segmentation onto this nested architecture creates a false dichotomy: researchers must either accept over-segmentation by treating every minor sub-topic shift as a major structural break, or suffer under-segmentation by blurring distinct subjects together under a single, coarse boundary.

The limitations of a purely linear approach become acutely apparent in noisy institutional environments. As observed in Chapter 4, while encoder-based models like SegNSP excel at detecting local topic shifts, they lack the internal scaffolding required to output the explicit, multi-level parent-child relationships that downstream retrieval systems demand. Furthermore, flat encoders remain highly susceptible to administrative boilerplate (e.g., adjournment formulas) that mimic topical shifts, resulting in severe structural fragmentation.

To overcome these limitations, this chapter presents **HSeg**, a two-stage generative **HTS** approach designed to extract nested semantic structures from unstructured prose. Leveraging the CitiLink-Minutes corpus [12] as a primary testbed, which provides a native two-level annotation

¹The source code, prompt templates, and evaluation scripts for the HSeg pipeline are publicly available at <https://github.com/jmisidro/hseg>.

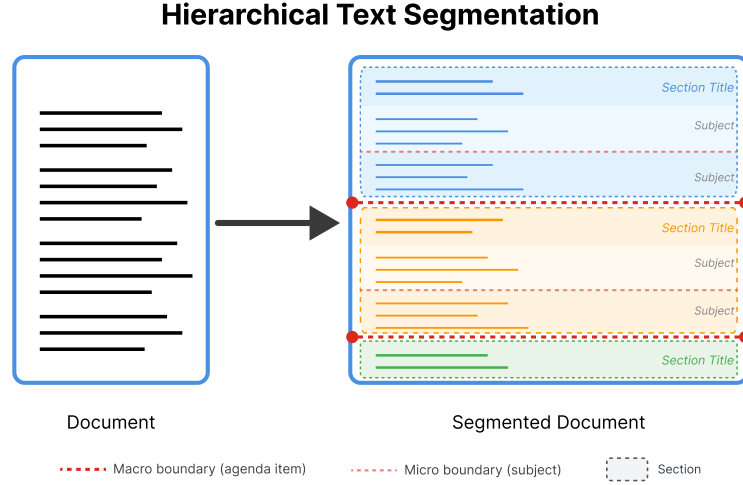


Figure 5.1: Illustration of **HTS**. A document D is partitioned into macro-level sections; each section is subsequently partitioned into micro-level subjects.

hierarchy comprising agenda items and constituent subjects, HSeg enforces a strict, top-down structural recovery. Figure 5.1 illustrates the resulting two-level decomposition.

The remainder of this chapter is structured as follows. Section 5.1 establishes the formal problem definition for hierarchical segmentation. Section 5.2 details the HSeg pipeline and its Sentence-ID grounding mechanics. Section 5.3 outlines the experimental setup, while Section 5.4 presents the empirical results alongside a prompt component ablation study. Finally, Section 5.5 synthesizes these findings across our core research questions.

5.1 Problem Formulation

Adopting a symmetrical formulation to the linear paradigm established in Section 4.1, let a document D be represented as an ordered sequence of n sentences produced by a standard sentence tokenizer:

$$D = (s_1, s_2, \dots, s_n).$$

The objective of **HTS** is to map this sequence into a discrete, two-level nested decomposition. At the macro level, D is partitioned into K contiguous, non-overlapping sections:

$$\mathcal{C} = (c_1, c_2, \dots, c_K),$$

where each section c_k is a contiguous sub-sequence of sentences:

$$c_k = (s_{p_k}, s_{p_k+1}, \dots, s_{q_k}), \quad p_k \leq q_k.$$

To ensure a valid macro-partition, the spans must be strictly exhaustive and non-overlapping: $p_1 = 1$, $q_K = n$, and $q_k + 1 = p_{k+1}$ for all $k < K$.

Subsequently, at the micro level, each macro-section c_k is further partitioned into J_k contiguous subjects:

$$\mathcal{S}_k = (\sigma_{k,1}, \sigma_{k,2}, \dots, \sigma_{k,J_k}),$$

where each subject $\sigma_{k,j}$ is a contiguous sub-sequence of sentences restricted entirely within the boundaries of its parent span c_k :

$$\sigma_{k,j} = (s_{u_{k,j}}, \dots, s_{v_{k,j}}), \quad p_k \leq u_{k,j} \leq v_{k,j} \leq q_k,$$

subject to the internal exhaustiveness constraints $u_{k,1} = p_k$, $v_{k,J_k} = q_k$, and $v_{k,j} + 1 = u_{k,j+1}$ for all $j < J_k$. The complete hierarchical structure for document D is therefore defined as the set of nested tuples:

$$\mathcal{H}(D) = \{(c_k, \mathcal{S}_k)\}_{k=1}^K.$$

5.2 Proposed Approach: The HSeg Pipeline

Transitioning from this mathematical formulation to a generative Large Language Model (LLM) paradigm introduces specific computational hurdles. A central challenge when applying generative LLMs to precise segmentation tasks is hallucination: models frequently alter the source text, fabricate boundaries, or generate summaries that cannot be deterministically mapped back to the original sequence. To mitigate this, HSeg relies on a strict Sentence-ID Grounding mechanism, a strategy inspired by the paragraph-wise identification approach utilized in LumberChunker [22]. Prior to any LLM inference, the document undergoes sentence tokenization, and each sentence s_i is assigned a unique identifier (e.g., $S_1, S_2, \dots, S_n$). The LLM is strictly prompted to express all boundary decisions using these identifiers rather than generating free text, reducing the degrees of freedom to a structured reference-selection task.

As illustrated in Figure 5.2, in Stage 1, the LLM processes the full, sentence-tokenized document to identify top-level sections (macro boundaries, $c_1 \dots c_K$). For each identified section c_k , the model is prompted recursively to locate nested sub-sections. This recursion continues dynamically until the model determines no further sub-sections exist within a given span, effectively identifying the leaf nodes of the discourse tree. Because this modular architecture treats every node uniformly, it effortlessly generalizes across datasets of varying depths. For instance, while municipal proceedings might naturally contain deeper nested sub-discussions, the CitiLink-Minutes dataset [12] is formally structured and evaluated as a two-level hierarchy. To align with this schema, the recursion can be constrained to a single macro-pass, preventing the LLM from hallucinating non-existent tertiary depths.

In Stage 2, the primary objective is to process each identified structural leaf node (section c_k) and extract its constituent micro-level subjects ($\sigma_{k,1} \dots \sigma_{k,J_k}$) by determining their precise sentence boundaries. In its standard configuration, the model focuses exclusively on retrieving these structural boundaries. While the standard HSeg pipeline functions independently using only boundary

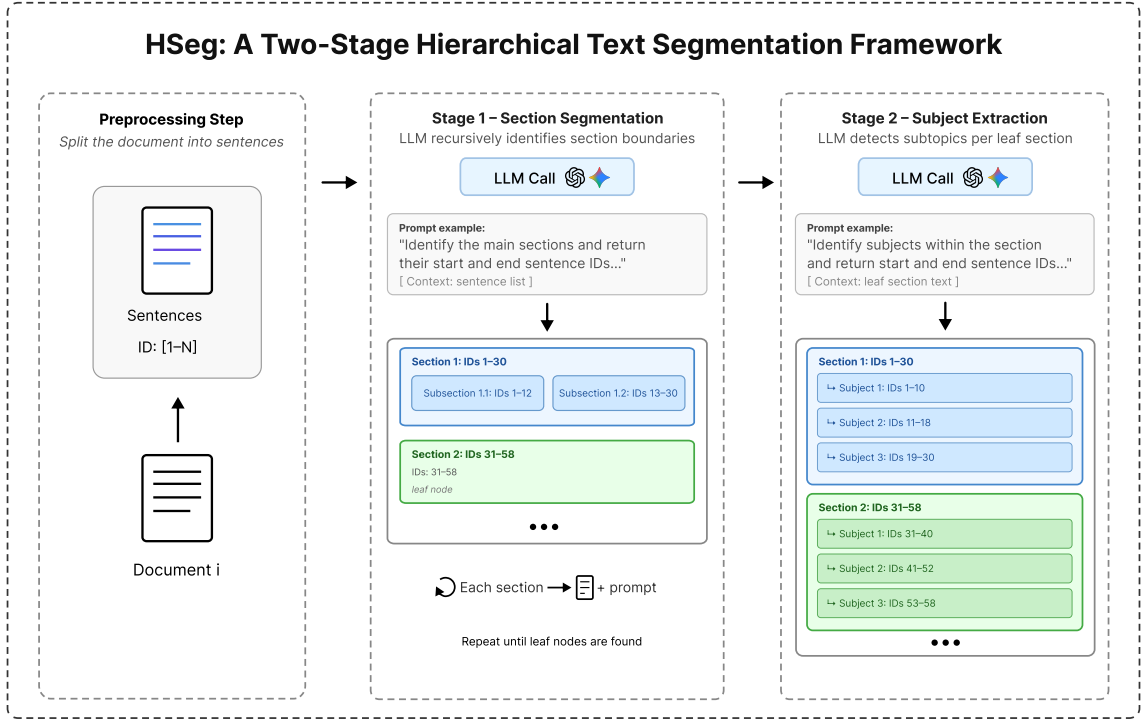


Figure 5.2: Overview of the HSeg segmentation pipeline.

detection, we hypothesize that this structural task can significantly benefit from explicitly engaging the model’s semantic understanding. Therefore, we introduce an optional configuration where the system is simultaneously tasked with generating a free-text theme $\tau_{k,j}$ and a multi-label topic classification $t_{k,j}$ for each subject $\sigma_{k,j}$. These classifications are drawn from a pre-defined, domain-specific taxonomy, which serves to standardize extracted subjects for downstream retrieval systems. We hypothesize that by forcing the model to categorize the text explicitly as it declares a boundary, this optional semantic labeling acts as an implicit reasoning scaffold, potentially improving the primary subject extraction task.

Operationally, successfully enforcing this sentence-grounding mechanism while executing the two-stage extraction process requires a highly structured prompt configuration to prevent schema drift. Rather than relying on static, free-form instructions, the HSeg pipeline dynamically compiles its generative queries at runtime using a standardized, four-module prompt anatomy:

1. **Domain Persona Assignment:** The model is explicitly assigned an authoritative persona (e.g., an expert municipal minutes analyst) via its system instructions. This serves to calibrate its internal logit distribution toward formal institutional prose.
2. **Enumerated Key-Value Injection:** The raw source text is passed into the context window not as standard paragraphs, but as an explicitly structured JSON-like object: $\{ID_i : s_i\}_{i=1}^n$. This forces the self-attention mechanism to explicitly bind semantic tokens to their prepended spatial coordinates.

3. **Dynamic Contextual Scaffolding:** When executing the micro-level extraction in Stage 2, the prompt dynamically ingests the parent agenda item’s declared title alongside localized municipal formatting rules derived during Stage 1. For instance, when processing records from the municipality of Porto, a specific `municipality_hint` is injected to warn the model to anticipate highly compressed, bulleted procedural outlines.
4. **Negative Constraint Enforcement:** To prevent the auto-regressive generation of conversational filler (e.g., “Here are the extracted subjects:”), the prompt terminates in a strict pseudo-schema. It explicitly lists negative boundary constraints, forbidding the model from synthesizing IDs outside the provided mapping array or breaking sub-items apart from their parent umbrellas.

To preserve the continuous narrative flow of this methodology, the complete, verbatim prompt templates utilized for both the CitiLink-Minutes and WIKI-727K extractions, alongside their full variable mappings, have been safely relegated to Appendix B.

5.3 Experimental Setup and Evaluation Framework

Datasets. To assess both domain-specific efficacy and open-domain generalization, experiments are evaluated across the following hierarchical datasets. Note that while the WikiSection corpus [3] was utilized in Chapter 4 to evaluate Linear Text Segmentation (LTS), it is intentionally excluded from this evaluation. Despite containing standardized section headings, WikiSection provides strictly flat, linear annotations and lacks the nested, multi-level ground truth necessary to evaluate a hierarchical framework. Therefore, we utilize the following datasets:

1. **CitiLink-Minutes [12]:** A domain-specific corpus comprising 120 city council meeting minutes from six Portuguese municipalities, featuring 2,885 manually annotated subjects nested within 1,403 agenda items. Crucially for our experiments, the corpus also provides ground-truth labels drawn from a 22-category municipal taxonomy (detailed at the end of Chapter 3 in Table 3.3). This allows us to empirically test the optional semantic generation configuration described in Section 5.2. Evaluation is conducted on the dataset’s default chronological test split of 24 documents, comprising the four most recent meeting minutes from each municipality.
2. **WIKI-727K [47]:** A large-scale, open-domain dataset containing 727,746 English Wikipedia documents. Based on the authors’ reported average of 3.48 segments per document, this yields over 2.5 million derived section annotations across the full corpus [47]. The dataset provides hierarchical structure derived automatically from tables of contents. The standard test set comprises roughly 73,000 documents. Due to the substantial financial and computational costs associated with proprietary LLM inference over extremely long contexts, we evaluate HSeg on a 10% uniform random sample of the test set (7,300 documents). This

sample size yields a margin of error of $\approx 1\%$ at a 95% confidence level, ensuring a statistically robust and credible approximation of the full test distribution.

3. **WIKI-50:** A lightweight, 50-document sample split derived from the WIKI-727K [47] test set, utilized to rapidly evaluate prompt architectures and establish baseline scaling metrics before executing large-volume inference.

Models Evaluated. We evaluated the proposed pipeline utilizing a spectrum of LLMs, systematically divided into proprietary and open architectures to ensure a comprehensive comparison:

- **Open-Weight and Open-Source Models:** We evaluated several recent instruction-tuned architectures. To explicitly test localized European language representations, we included the fully open-source Assistente Multimodal Automático de Linguagem com Inteligência Artificial (**AMALIA**) (9B) [78], alongside the multilingual EuroLLM-9B-Instruct and EuroLLM-22B-Instruct models [59]. Additionally, we benchmarked general open-weight frontier models, including Qwen2.5-7B-Instruct [89], Ministral-8B-Instruct-2410 [1], and Gemma-4-12B-it [31].
- **Proprietary Models:** To benchmark against frontier models, we utilized Gemini 2.5 Flash Lite [18], Gemini 3.1 Flash Lite [37], Claude Opus 4.7 [2], and GPT-5.5 [69].

Inference and Decoding Parameters. To ensure high structural adherence and suppress autoregressive stochasticity, inference across all evaluated generative models was executed using near-greedy, low-temperature decoding (temperature = 0.1, top_p = 1.0). Constraining the distribution to a 0.1 temperature is vital for pointer-based segmentation; it minimizes coordinate drift while granting the network a marginal degree of lexical flexibility when synthesizing the reasoning scaffolds.

For the open-domain evaluation on WIKI-727K, HSeg is compared against an array of established sequence-labeling and neural models sourced from the literature. Specifically, we report the performance of **Bi-LSTM** [47], **Cross Segment BERT**, **BERT+Bi-LSTM**, and **Hierarchical BERT** [56]². Additionally, we include state-of-the-art CRF-based models, **HierCRF-BERT** and **HierCRF-Aug-BERT** [62], as well as the recent **MiniSeg** architecture [74].

Metrics. Boundary detection across both datasets is assessed using a comprehensive suite of metrics. For the WIKI-727K open-domain evaluation, we expand our evaluation to report Precision (P), Recall (R), and F₁ Score, as these are the standard metrics used to evaluate the WIKI-727K benchmark in the literature. We report these alongside P_k , Boundary Similarity (B), and Boundary F _{β} -measure ($B-F_1$). For CitiLink-Minutes, we primarily report $B-F_1$, B , P_k , and WindowDiff (WD) using a tolerance window of two sentences to accommodate subtle administrative topic shifts. Baseline results for WIKI-727K are sourced directly from their original publications, and metrics not provided by the respective authors are denoted with a dash.

²Results for these four baselines are reported as by Lukasik et al. [56].

Ablation Study Configuration. To explore whether the optional semantic generation capability introduced in Section 5.2 provides any structural benefit, we conducted an auxiliary ablation study on the CitiLink-Minutes test split. It is important to note that this study is exploratory in nature: the Full Schema configuration was *not* used in the primary multi-model evaluation reported in Tables 5.1 and 5.2, nor in the WIKI-727K experiments. The canonical HSeg result throughout this dissertation is therefore the *Boundaries Only* configuration, which represents the standard pipeline. This ablation is designed solely to quantify the potential upper-bound benefit of semantic scaffolding. We evaluate the frontier Claude Opus 4.7 model under three prompt conditions: *Boundaries Only* (the standard pipeline), *Boundaries + Theme* (additionally requiring a free-text summary), and *Boundaries + Theme + Topic* (requiring both a title and a classification drawn from the municipal taxonomy, detailed at the end of Chapter 3 in Table 3.3). To ensure the generated reasoning scaffolds remain semantically faithful to the source text, we report the Theme BERTScore F1 and Topic F1 alongside the standard structural boundary metrics.

To explore whether the optional semantic generation capability introduced in Section 5.2 provides any structural benefit, we conducted an auxiliary ablation study on the CitiLink-Minutes test split. It is important to note that this study is exploratory in nature: the Full Schema configuration was *not* used in the primary multi-model evaluation reported in Tables 5.1 and 5.2, nor in the WIKI-727K experiments. The Full Schema requires the model to generate additional free-text and classification output per subject, increasing both output token consumption and the input context of subsequent recursive stages, which raises inference cost at evaluation scale. The canonical HSeg result throughout this dissertation is therefore the *Boundaries Only* configuration ($B-F_1 = 89.16$), which represents the standard, cost-effective pipeline. This ablation is designed solely to quantify the potential upper-bound benefit of semantic scaffolding, should future work or reduced API costs make the Full Schema operationally viable. We evaluate the frontier Claude Opus 4.7 model under three prompt conditions: *Boundaries Only* (the standard pipeline), *Boundaries + Theme* (additionally requiring a free-text summary), and *Boundaries + Theme + Topic* (additionally requiring a classification from the municipal taxonomy in Table 3.3). Theme BERTScore F1 and Topic F1 are reported alongside the structural metrics to verify semantic faithfulness.

5.4 Results

This section presents the empirical evaluation of the HSeg pipeline. It details the segmentation performance across both the domain-specific administrative corpus (CitiLink-Minutes) and the open-domain Wikipedia benchmark (WIKI-727K). Furthermore, it includes an ablation study designed to isolate the impact of semantic generation as an implicit reasoning scaffold during the structural extraction process.

5.4.1 Domain-Specific Evaluation (CitiLink-Minutes)

To evaluate the efficacy of the proposed pipeline on administrative texts, we first examine the models’ capacity to construct the structural skeleton of the documents. Table 5.1 reports the macro-level boundary detection performance for identifying top-level agenda items.

Table 5.1: Stage 1 section segmentation results on the CitiLink-Minutes test split. Higher $B-F_1/B$ and lower P_k/WD indicate better performance.

Architecture	Model	$B-F_1 \uparrow$	$B \uparrow$	$P_k \downarrow$	$WD \downarrow$
<i>HSeg (Proprietary)</i>	Claude Opus 4.7	80.18	47.13	20.02	35.03
	GPT-5.5	69.46	41.09	28.88	44.44
	Gemini 3.1 Flash Lite	55.50	38.16	29.51	28.12
	Gemini 2.5 Flash Lite	58.19	38.76	31.75	38.01
<i>HSeg (Open)</i>	Ministral-8B-Instruct-2410	40.39	25.74	41.21	51.92
	Qwen2.5-7B-Instruct	31.28	21.41	40.45	48.16
	Gemma-4-12B-it	23.19	11.45	24.46	28.12
	EuroLLM-22B-Instruct	21.16	13.24	44.67	51.74
	EuroLLM-9B-Instruct	13.60	8.05	40.29	44.73
	AMALIA (9B)	12.75	8.06	38.11	43.28

At the macro-level, frontier proprietary models demonstrated a decisive advantage in navigating complex administrative boilerplate. Claude Opus 4.7 established the strongest structural scaffold, achieving a $B-F_1$ of 80.18. By contrast, the open architectures experienced near-total systemic failure. Despite evaluating models with strong instruction-tuning capabilities (Ministral-8B, Qwen2.5) and localized European Portuguese pre-training (**AMALIA**), none were able to reliably extract the macro-structure. Ministral-8B achieved the highest open-weight $B-F_1$ at merely 40.39, while **AMALIA** struggled significantly to capture valid boundaries ($B-F_1$ of 12.75). This stark contrast suggests that parsing noisy, long-context municipal minutes demands a degree of zero-shot reasoning and context-window resilience that mid-sized open-weight/open-source models currently lack, causing them to hallucinate boundaries or ignore procedural shifts entirely.

Crucially, benchmarking the EuroLLM family at this macro-layer isolates the operational impact of raw parameter scaling within localized European models. While EuroLLM-9B exhibited severe structural limitations similar to **AMALIA** (achieving a macro $B-F_1$ of 13.60), scaling the architecture up to 22 billion parameters (EuroLLM-22B) nearly doubled this retention to 21.16. However, because an absolute score of 21.16 still represents severe structural failure, it demonstrates that parameter bulk alone cannot compensate for a lack of specialized long-context procedural training when parsing the complex boilerplate of municipal agendas.

This macro-level accuracy directly dictates the success of the entire pipeline, as the identified spans provide the contextual limits for subsequent subject extraction. Building upon these macro-level foundations, Table 5.2 details the micro-level (subject) boundary detection performance of the HSeg pipeline.

Table 5.2: Stage 2 (Subject level) boundary detection results evaluated within the HSeg pipeline framework.

Architecture	Model	$B\text{-}F_1 \uparrow$	$B \uparrow$	$P_k \downarrow$	WD $\downarrow$
<i>HSeg (Proprietary)</i>	Claude Opus 4.7	89.16	62.33	7.23	21.89
	GPT-5.5	84.61	58.58	9.06	24.42
	Gemini 3.1 Flash Lite	81.01	62.16	8.61	18.57
	Gemini 2.5 Flash Lite	62.93	49.98	20.63	34.52
<i>HSeg (Open)</i>	Qwen2.5-7B-Instruct	37.55	30.07	33.64	48.77
	Ministral-8B-Instruct-2410	36.26	28.33	37.55	55.12
	EuroLLM-22B-Instruct	33.45	20.37	39.89	53.80
	Gemma-4-12B-it	25.81	14.73	33.31	36.98
	EuroLLM-9B-Instruct	20.06	11.70	43.04	50.38
	AMALIA (9B)	13.39	13.32	44.74	54.90

The incorporation of hierarchical structural information significantly improves micro-level segmentation, but only when the underlying architecture possesses the necessary reasoning capacity. When utilizing proprietary frontier models such as Claude Opus 4.7, the hierarchical scaffold successfully guides the model to place micro-level boundaries with exceptional precision, yielding a state-of-the-art subject-level $B\text{-}F_1$ of 89.16 and a remarkably low P_k of 7.23.

Conversely, the open models tier suffered heavily from error propagation. Because these architectures fundamentally failed to establish correct macro-level boundaries during Stage 1, the text spans fed into Stage 2 were inherently flawed. This resulted in cascading structural fragmentation, with Qwen2.5-7B topping the open-weight micro-level results at a $B\text{-}F_1$ of 37.55. The high discrepancy in WindowDiff scores across these models confirms widespread structural misinterpretation rather than minor sentence-level boundary misses.

This scaling dynamic is mirrored within the localized European models at the micro-level. While EuroLLM-9B struggled severely to partition individual deliberations ($B\text{-}F_1$ of 20.06), EuroLLM-22B achieved a measurable leap to 33.45. Yet, despite this +66% relative performance gain over its 9B sibling, EuroLLM-22B remains unable to surpass the localized exactitude of the much smaller, highly instruction-tuned Qwen2.5-7B model (37.55). This suggests that when navigating the dense, noisy boundaries of administrative subjects, a model’s foundational instruction-following adherence dictates structural recovery far more than raw parameter capacity.

5.4.2 Open-Domain Evaluation (WIKI-727K)

To validate the generalizability of HSeg beyond administrative boilerplate, we applied the recursive adaptation of our pipeline to the WIKI-727K dataset. The evaluation schema here necessarily diverges from the two-table structure (macro-level and micro-level) utilized for the CitiLink-Minutes corpus.

Unlike CitiLink-Minutes, which possesses a strict two-level annotation schema (macro agenda items and micro subjects), WIKI-727K features a variable-depth, n-ary hierarchical structure derived from Wikipedia tables of contents. In standard open-domain document segmentation research, this hierarchy is flattened into a single linear sequence, treating all headings and subheadings, regardless of their nested depth, as uniform segment boundaries [47].

Consequently, rather than separating the evaluation into discrete hierarchical tiers, we evaluate the final output of the HSeg pipeline. To ensure a direct and fair comparison with the literature baselines, we report the leaf-level performance. Because these micro-level sub-sections partition the entire document end-to-end without gaps, their boundaries naturally represent the complete union of all macro-level and nested transition points. This approach perfectly aligns with the flattened ground truth utilized in prior work, allowing for an accurate calculation of the standard sentence-level metrics.

Table 5.3: Hierarchical boundary detection results on the WIKI-727K test set (10% random sample for HSeg models). Baseline metrics are sourced directly from their respective publications. Higher Precision (P), Recall (R), F_1 , $B-F_1$, and B indicate better performance, while lower P_k is preferred. [†]Results as reported by Lukasik et al. [56].

Architecture	Model	P $\uparrow$	R $\uparrow$	F_1 $\uparrow$	$B-F_1$ $\uparrow$	P_k $\downarrow$	B $\uparrow$
<i>Baselines</i>	Bi-LSTM [†] [47]	69.30	49.50	57.70	[–]	[–]	[–]
	Cross Segment BERT [†] [56]	69.10	63.20	66.00	[–]	[–]	[–]
	BERT + Bi-LSTM [†] [56]	67.30	53.90	59.90	[–]	[–]	[–]
	Hierarchical BERT [†] [56]	69.80	63.50	66.50	[–]	[–]	[–]
	HierCRF-BERT [62]	79.00	63.30	70.20	[–]	[–]	[–]
	HierCRF-Aug-BERT [62]	80.40	64.60	71.60	[–]	[–]	[–]
	MiniSeg [74]	68.57	70.76	69.65	[–]	17.57	59.81
<i>HSeg (Proprietary)</i>	Claude Opus 4.7	86.81	72.84	78.83	89.26	5.30	81.88
	GPT-5.5	85.89	71.34	77.61	88.55	5.81	81.03
	Gemini 3.1 Flash Lite	83.62	79.51	80.66	90.14	7.30	81.30
	Gemini 2.5 Flash Lite	78.65	74.88	75.39	85.16	9.77	75.93
<i>HSeg (Open)</i>	Gemma-4-12B-it	79.85	75.80	76.65	87.14	7.94	77.97
	EuroLLM-22B-Instruct	48.56	39.86	42.61	48.81	30.19	34.47
	Qwen2.5-7B-Instruct	35.16	55.43	40.30	49.49	35.41	35.33
	Ministral-8B-Instruct-2410	26.23	54.59	31.94	40.15	38.57	26.47
	EuroLLM-9B-Instruct	19.37	55.22	25.45	32.48	47.57	20.28
	AMALIA (9B)	15.31	38.13	19.68	26.02	49.81	15.63

As detailed in Table 5.3, within the evaluated representative sample of the WIKI-727K benchmark, the recursive HSeg pipeline achieves what constitutes, to the best of our knowledge, state-of-the-art performance among dedicated segmentation architectures. While these figures are strictly bounded by our sampling margin of error ($\pm 1\%$), the absolute performance gains are pronounced.

The highly optimized Gemini 3.1 Flash Lite established the highest overall harmonic balance, yielding an F_1 score of 80.66. This surpasses the previous non-generative baseline (HierCRF-Aug-BERT) by an absolute margin of +9.06 points, and outperforms even the much larger frontier

Claude Opus 4.7 model (78.83). Conversely, Claude Opus 4.7 retained the lead in strict exactitude, achieving the highest leaf Precision (86.81) and the lowest probabilistic penalty ($P_k = 5.30$). This dichotomy reveals that while massive frontier models excel at precision and minimizing probabilistic penalty gaps, lightweight generative models like Gemini 3.1 can deliver unmatched recall and overall F-measure balance.

Scalability and Robustness Across Dataset Sizes. To further validate the robustness of the HSeg pipeline and confirm that our 10% sampling strategy does not misrepresent the model’s capabilities, we compared performance across two distinct scales: the WIKI-50 pilot split and the larger WIKI-727K test sample.

Table 5.4: Comparison of HSeg performance across dataset scales (WIKI-50 vs. WIKI-727K). Results indicate stable generalization capabilities irrespective of inference volume.

Dataset	Model	P ↑	R ↑	F ₁ ↑	B-F ₁ ↑	P _k ↓	B ↑
WIKI-50	<i>Proprietary</i>						
	Claude Opus 4.7	82.36	70.85	75.90	84.56	6.57	78.91
	GPT-5.5	82.03	67.85	74.09	85.95	7.28	78.18
	Gemini 3.1 Flash Lite	77.63	77.52	76.41	86.64	8.97	78.37
	Gemini 2.5 Flash Lite	70.71	72.51	69.61	82.37	11.33	72.83
	<i>Open</i>						
	Gemma-4-12B-it	78.87	79.02	77.30	88.65	9.49	80.74
	EuroLLM-22B-Instruct	45.89	39.53	41.31	40.73	35.49	26.08
	Qwen2.5-7B-Instruct	31.70	52.60	38.14	46.18	37.55	31.52
	Ministral-8B-Instruct-2410	29.72	55.55	33.63	44.16	38.89	31.12
	EuroLLM-9B-Instruct	16.13	58.54	20.15	26.17	43.44	15.14
	AMALIA (9B)	16.09	40.03	20.21	27.24	45.46	17.24
WIKI-727K	<i>Proprietary</i>						
	Claude Opus 4.7	86.81	72.84	78.83	89.26	5.30	81.88
	GPT-5.5	85.89	71.34	77.61	88.55	5.81	81.03
	Gemini 3.1 Flash Lite	83.62	79.51	80.66	90.14	7.30	81.30
	Gemini 2.5 Flash Lite	78.65	74.88	75.39	85.16	9.77	75.93
	<i>Open</i>						
	Gemma-4-12B-it	79.85	75.80	76.65	87.14	7.94	77.97
	EuroLLM-22B-Instruct	48.56	39.86	42.61	48.81	30.19	34.47
	Qwen2.5-7B-Instruct	35.16	55.43	40.30	49.49	35.41	35.33
	Ministral-8B-Instruct-2410	26.23	54.59	31.94	40.15	38.57	26.47
	EuroLLM-9B-Instruct	19.37	55.22	25.45	32.48	47.57	20.28
	AMALIA (9B)	15.31	38.13	19.68	26.02	49.81	15.63

As detailed in Table 5.4, the pipeline maintains stable performance across all evaluated metrics regardless of the inference volume. Notably, the models exhibited stability or significant improvements when scaling up to the 7,300-document test set. Gemini 3.1 Flash Lite, in particular, demonstrated a massive leap in performance at scale, climbing from an F₁ score of 76.41 on

the pilot split to 80.66 on the large sample, alongside an unprecedented $B-F_1$ of 90.14.

Equally remarkable is the performance of the open-weight Gemma-4-12B-it model on the WIKI-50 pilot split. It achieved a F_1 score of 77.30 and a $B-F_1$ of 88.65, effectively surpassing the proprietary frontier Claude Opus 4.7 model (F_1 of 75.90) in this smaller inference volume. Although Claude Opus 4.7 reclaimed the lead at scale on the WIKI-727K set, Gemma-4-12B-it’s ability to punch well above its weight class reinforces the exceptional viability of highly optimized mid-sized open-weight architectures.

This open-domain environment also confirms an internal scaling law within the EuroLLM family: transitioning from EuroLLM-9B to EuroLLM-22B yields a +17.16 absolute point jump in F_1 score (25.45 vs. 42.61), allowing the 22B model to comfortably surpass Qwen2.5-7B and Ministral-8B. Yet, the stark reality remains that EuroLLM-22B is still outperformed by the much smaller Gemma-4-12B model by a massive 34.04 F_1 points. This reinforces a central takeaway of our structural evaluations: for complex, pointer-based document structuring, an open model’s post-training alignment quality and structured instruction-following capabilities carry vastly more weight than raw parameter bulk.

Nevertheless, the tight variance and upward scaling trajectory across both proprietary and open-weight models confirm that the HSeg architecture does not overfit to small samples and seamlessly generalizes to massive open-domain text corpora, cementing the validity of our methodological sampling approach.

5.4.3 Semantic Schema as a Reasoning Scaffold

Table 5.5: Ablation study isolating the impact of semantic metadata generation on micro-level boundary detection accuracy (CitiLink-Minutes test split, Claude Opus 4.7).

Prompt Condition	$B-F_1 \uparrow$	Theme BERTScore $F1 \uparrow$	Topic $F1 \uparrow$
Full Schema (<i>Boundaries + Theme + Topic</i>)	90.08	76.83	74.20
w/o Topic Classification (<i>Boundaries + Theme</i>)	89.90	76.81	[–]
w/o Semantic Scaffold (<i>Boundaries Only</i>)	89.16	[–]	[–]

As introduced in Section 5.2, we hypothesized that explicitly tasking the model with generating semantic metadata alongside its boundary pointers would engage an implicit chain-of-thought that benefits structural precision. To test this, we evaluated Claude Opus 4.7 under three prompt conditions on the CitiLink-Minutes test split, progressively removing semantic generation requirements.

The empirical trajectory, detailed in Table 5.5, supports our hypothesis. Operating the model under the *Full Schema* yields the highest observed structural accuracy ($B-F_1 = 90.08$). Ablating the requirement to classify the text against the municipal taxonomy (*w/o Topic Classification*) results in a minor degradation to 89.90. Further ablating the free-text summarization requirement, reverting the prompt to the purely structural *Boundaries Only* baseline, performance drops to 89.16, which is the canonical HSeg result adopted throughout this dissertation. This stepwise

trajectory suggests that articulating the semantic core of a span prior to declaring boundary coordinates provides a small but consistent benefit (+0.92 B- F_1 points over the standard pipeline).

However, the *Full Schema* increases output token consumption and input context size per subject, raising inference cost at scale, which precluded its use in the primary multi-model evaluation. It is therefore best understood as an exploratory upper-bound estimate, demonstrating the potential gains achievable should inference costs allow a full multi-model evaluation under this configuration in future work.

5.5 Discussion

The empirical evaluations across both domain-specific and open-domain corpora allow us to synthesize several key insights and formally address our core research questions.

First, these findings directly address **RQ4**. By transitioning from the flat linear boundary detection of models like SegNSP (which achieved a B- F_1 of 79.22 on our adaptation of CitiLink-Minutes) to the top-down, two-stage structural recovery of HSeg, performance on the administrative domain elevated to a state-of-the-art micro-level B- F_1 of 89.16. Linear models inherently flatten discourse, which frequently causes them to confuse administrative boilerplate or nested sub-discussions for main procedural shifts. HSeg overcomes this because the macro-level scaffold dictates strict contextual limits for the subsequent micro-level extraction, proving that explicit hierarchical modeling significantly curbs both over- and under-segmentation.

Furthermore, the results explicitly answer **RQ3**. The empirical evidence demonstrates that generative **LLMs**, when deployed in a structured hierarchical framework, surpass established supervised Transformer baselines. While traditional encoder-based models such as HierCRF-Aug-BERT and MiniSeg remain computationally efficient for flat sequence labeling, the advanced reasoning and contextual semantic understanding of frontier generative **LLMs** allow them to establish a new State-of-the-Art (**SotA**) in detecting variable-depth topic boundaries. This is evidenced by the +9.06 absolute F_1 margin over prior dedicated architectures on the WIKI-727K benchmark.

Finally, synthesizing the proprietary and open-weight performances across both datasets highlights a critical caveat regarding current open-weight deployment for civic technology. The evaluation demonstrates that the architectural advantage of HSeg is not strictly gated behind massive proprietary compute, but generalization remains highly dependent on a model’s pre-training distribution. The open-weights Gemma-4-12B-it model demonstrated exceptional robustness on standardized English encyclopedic text, successfully outperforming all traditional CRF-based baselines without relying on a paid API. Conversely, it lacked the cross-lingual reasoning required to process administrative Portuguese text.

In stark contrast, models like **AMALIA** (9B) and Ministral-8B struggled significantly outside their pre-training distributions. While models like **AMALIA** possess a strong foundational understanding of European Portuguese, they suffered from catastrophic generalization failure when

tasked with zero-shot structural extraction on open-domain English data (F_1 of 19.68 on WIKI-727K). Furthermore, they still lacked the reasoning capacity and context-window resilience required to process the dense municipal minutes. Consequently, until parameter-efficient open-weight models are explicitly fine-tuned for administrative segmentation, frontier proprietary APIs currently represent the most viable and reliable out-of-the-box solution for executing zero-shot hierarchical extraction in noisy, real-world domains.

5.6 Summary

This chapter introduced HSeg, a recursive, generative framework designed to tackle the deeply nested discourse structures of complex texts. Transitioning from the flat linear boundary detection explored previously, HSeg formulates Hierarchical Text Segmentation (**HTS**) as a structured two-stage extraction process. To mitigate the generative hallucinations commonly associated with **LLMs**, the pipeline employs a strict Sentence-ID grounding mechanism. Furthermore, through an ablation study on prompt semantics, we demonstrated that explicitly tasking the model with generating descriptive metadata, such as themes and domain-specific topics, acts as a highly effective implicit reasoning scaffold, slightly improving structural boundary placement.

Empirical evaluations across diverse benchmarks confirmed the efficacy of this approach. On the domain-specific CitiLink-Minutes dataset, frontier proprietary **LLMs** successfully mapped complex administrative macro- and micro-structures, achieving a high micro-level B- F_1 of 89.16. Concurrently, the framework accommodated the integration of localized open-weight architectures like **AMALIA**, highlighting viable, parameter-efficient alternatives for institutional processing. When applied to the open-domain WIKI-727K benchmark, the recursive HSeg architecture established a new overall State-of-the-Art (**SotA**). Highly optimized lightweight models surpassed previous dedicated sequence-labeling architectures by an absolute margin of +9.06 F_1 points, while frontier models minimized probabilistic error rates to unprecedented levels.

Ultimately, the findings presented in this chapter address a critical gap in the field of text segmentation. While traditional linear sequence-labeling models remain computationally efficient, generative **LLMs** equipped with targeted structural prompting and implicit semantic reasoning provide superior adaptability, effectively capturing the multi-level organization of real-world documents. Building upon the foundational datasets, linear baselines, and hierarchical frameworks established throughout this work, the following chapter demonstrates the practical deployment of these models into interactive civic technology platforms before the dissertation concludes.

Chapter 6

Practical Applications and Demonstrations

The preceding chapters detailed the theoretical foundations, architectural designs, and empirical evaluations of the proposed Text Segmentation (TS) models. Chapter 4 introduced SegNSP, a linear sequence-labeling framework tailored for Linear Text Segmentation (LTS), while Chapter 5 presented HSeg, a generative approach for Hierarchical Text Segmentation (HTS). While intrinsic quantitative metrics such as Boundary F_β -measure ($B-F_1$) and P_k are essential for benchmarking these models against established baselines, they do not fully capture the practical utility, error topologies, and qualitative nuances of the segmentation outputs.

In applied Natural Language Processing (NLP) and Information Retrieval (IR), bridging the gap between empirical evaluation and real-world deployment is critical. This chapter presents a series of interactive demonstrations and user interfaces developed alongside the core models. These tools were engineered to facilitate the human-in-the-loop evaluation of segmentation outputs, enable visual error analysis, and ultimately integrate the underlying NLP pipelines into a production-ready civic technology platform¹.

The remainder of this chapter is structured as follows. Section 6.1 presents the interactive web application developed to visualize and debug the SegNSP linear segmentation process. Section 6.2 details the diagnostic dashboard built for the qualitative and visual evaluation of the generative HSeg pipeline. Section 6.3 demonstrates the extrinsic application of these models, detailing their end-to-end integration into the CitiLink civic technology platform.

6.1 Interactive Linear Segmentation: The SegNSP Demo

To make the LTS process tangible and inspectable, a web-based demonstration was developed for the SegNSP model². Hosted as an interactive application, this interface allows researchers and

¹<https://citilink.inesctec.pt/>

²<https://huggingface.co/inesctec/CitiLink-NSP-Segmentation-pt>

end-users to input raw administrative text, such as city council minutes, and instantly visualize the model’s structural predictions.

The interface is built around transparency and user control. A core feature of the SegNSP architecture is its reliance on a probabilistic decision threshold to determine whether a sentence pair constitutes a topic shift. In the demo, this threshold is exposed via an interactive slider. As the user adjusts the threshold (e.g., setting it to 0.50), the application dynamically recalculates the boundaries and visually separates the text into distinct, color-coded segments.



Figure 6.1: The SegNSP text segmentation demonstration interface, showcasing the interactive threshold slider and real-time document segmentation visualization.

Alongside the segmented document view, the interface provides immediate summary statistics, including the total number of extracted segments, total sentences processed, and the average number of sentences per segment (Figure 6.1). This real-time feedback loop proved invaluable during the development phase, which allowed us to qualitatively assess the model’s sensitivity to administrative boilerplate and to visually identify instances of over- or under-segmentation that aggregate window-based metrics might obscure.

Furthermore, the interface explicitly visualizes the model’s confidence at each transition point (Figure 6.2). By rendering the Next Sentence Prediction (NSP) boundary detection probability alongside the resulting text blocks, researchers can conduct granular error analysis, observing exactly how the model negotiates ambiguous discourse markers and where the continuity signal begins to decay.

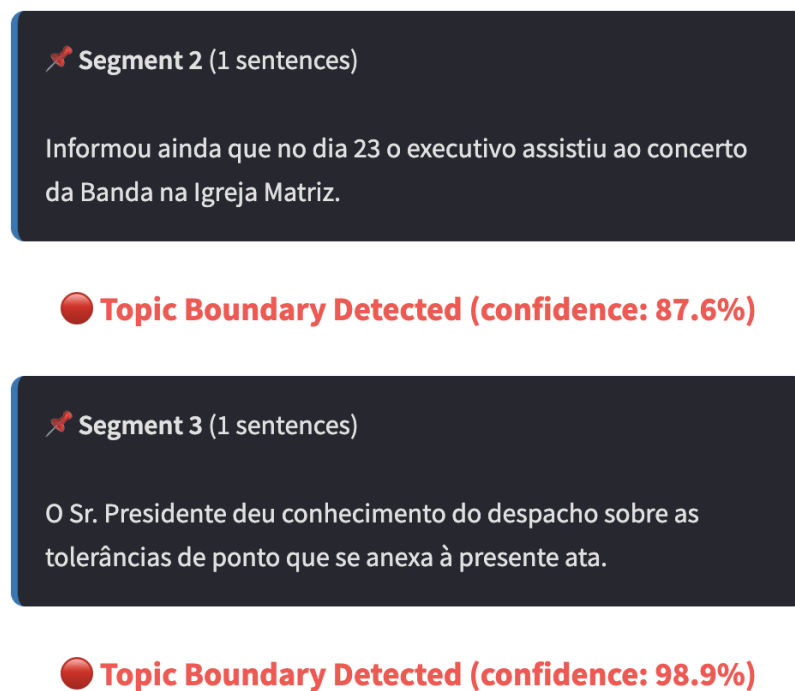


Figure 6.2: Detailed view of the SegNSP output, explicitly rendering the topic boundary detection confidence scores between extracted segments to aid in error analysis.

6.2 Qualitative Analysis and Model Comparison: HSeg Evaluation Interface

Evaluating the complex, multi-level outputs of generative Large Language Models (LLMs) requires moving beyond flat metric tables. To support the rigorous evaluation detailed in Chapter 5, a dedicated Streamlit-based diagnostic dashboard was developed for our HSeg framework³. This evaluation interface was specifically designed to handle the recursive outputs of the HSeg pipeline, mapping predicted macro-boundaries (Agenda Items) and micro-boundaries (Subjects) against the ground truth annotations.

The dashboard consists of several specialized analytical modules designed to surface distinct types of evaluation insights:

- **Per-Document Distributions:** This view allows for a granular inspection of performance across different municipalities (Figure 6.3). By breaking down metrics on a per-document basis, we can immediately identify which local procedural formatting conventions cause the generative models to struggle, isolating out-of-distribution stylistic shifts.
- **Macro-Level Alignment (Agenda Items):** This structural view renders a direct, side-by-side visual alignment of the ground truth macro-segments versus the predicted Agenda Items

³https://huggingface.co/spaces/jmisidro/hseg_demo

(Figure 6.4). By visually juxtaposing these structural blocks, we can instantly detect where the generative model successfully parsed broad procedural umbrellas and where it fabricated or omitted major transitions. This immediate visual feedback significantly accelerates the human-in-the-loop evaluation process, bypassing abstract aggregate metrics to allow for intuitive, block-level structural debugging.

- **Micro-Level Alignment (Subjects):** Complementing the macro-view, the dashboard provides a deeper, side-by-side visual inspection of the nested subjects (Figure 6.5). This granular visualization is crucial for identifying cascading errors, such as when a misaligned macro-boundary inherently skews the context window for the subsequent subject extraction stage. By visually comparing the predicted segment bounds, generated themes, and taxonomy categories directly against the ground truth annotations, we can rapidly assess the qualitative coherence of the generated hierarchy, ensuring that the model’s structural reasoning aligns with human intuition.
- **Run Comparison:** To facilitate the benchmarking of different generative architectures (e.g., Claude Opus 4.7, GPT-5.5, and Gemini 2.5 Flash Lite), this module plots comparative bar charts for metrics such as P_k , Boundary Similarity (B), and Boundary F_β -measure ($B-F_1$) (Figure 6.6). This visual representation quickly communicates the precision-recall trade-offs between frontier proprietary models and highly compressed local alternatives.

 **Per-Document Metric Distribution**

Document	Municipality	AI Boundary F1	AI BED F-measure	Subj Boundary F1	Subj BED F-measure	Subj Boundary Sim	AI #Predicted	AI #GT	Subj #Predicted	Subj #GT
Alandroal_cm_011_2024-04-24	Alandroal	0.9375	0.625	1	1	0.9667	16	16	17	17
Alandroal_cm_015_2024-06-19	Alandroal	0.9375	0.6458	0.9756	0.9552	0.8889	16	16	20	21
Alandroal_cm_019_2024-08-07	Alandroal	0.9	0.6491	0.898	0.975	0.907	20	20	22	27
Alandroal_cm_021_2024-09-11	Alandroal	0.9	0.6393	0.902	0.9639	0.9091	20	20	23	28
Campomaor_cm_007_2024-04-03	Campomaor	1	0.6154	0.9167	0.6774	0.5	6	6	12	12
Campomaor_cm_010_2024-05-15	Campomaor	0.6667	0.4444	1	0.6667	0.4889	4	5	24	24
Campomaor_cm_018_2024-10-02	Campomaor	1	0.6	1	0.6889	0.5	5	5	18	18
Campomaor_cm_020_2024-11-06	Campomaor	0.6	0.6	0.878	0.6545	0.4737	5	5	21	20
Covilha_cm_006_2024-03-28	Covilha	1	0.6667	1	0.7742	0.5714	13	13	22	22
Covilha_cm_010_2024-06-17	Covilha	1	0.7143	1	0.6812	0.5	14	14	22	22

Figure 6.3: The Per-Document Metric Distribution view, allowing researchers to isolate model performance by specific municipal domains.

By providing a visual medium for structural error analysis, this diagnostic tool was instrumental in refining the prompt architectures and strict Sentence-ID grounding mechanisms utilized in the final HSeg framework.

6.3 Production Deployment: The CitiLink Platform

The ultimate objective of developing these segmentation models is to enhance the accessibility and navigability of public municipal records. To demonstrate this end-to-end value, the segmen-

Agenda Item Alignment

Ground Truth

1. INFORMAÇÕES
 General Administration, Finance, and Human Resources Culture
 Communication and Public Relations
 • Informação sobre o relatório da ROC em relação à prestação de contas de 2023
 • Participação do executivo na apresentação do Festival Soil to Soul
 IoU: 0.988

2. ALTERAÇÃO ORÇAMENTAL PERMUTATIVA PARA RATIFICAÇÃO
 General Administration, Finance, and Human Resources
 • Alterações orçamentais permutativas
 IoU: 0.987

3. ALTERAÇÃO ORÇAMENTAL PERMUTATIVA
 General Administration, Finance, and Human Resources
 • Ausência de alteração orçamental permutativa
 IoU: 0.960

4. APROVAÇÃO DE ATA
 General Administration, Finance, and Human Resources
 • Aprovação da ata n.º 9, de 10.04.2024
 IoU: 0.967

Predicted

1. INFORMAÇÕES
 1. INFORMAÇÕES
 O Sr. Presidente deu conhecimento do relatório da ROC em relação à Prestação de contas de 2023. Continuou por informar que no dia 19 o executivo participou na apresentação do Festival S...
 2 subject(s) extracted
 IoU: 0.988

2. ALTERAÇÃO ORÇAMENTAL PERMUTATIVA PARA RATIFICAÇÃO
 ALTERAÇÃO ORÇAMENTAL PERMUTATIVA PARA RATIFICAÇÃO
 Pelo Sr. Presidente foram presentes a reunião as informações da contabilidade que se anexam à presente ata.
 Ponderado e analisado o assunto o Executi...
 1 subject(s) extracted
 IoU: 0.987

3. ALTERAÇÃO ORÇAMENTAL PERMUTATIVA
 3. ALTERAÇÃO ORÇAMENTAL PERMUTATIVA
 Não houve alteração orçamental permutativa para apresentar.
 4.
 1 subject(s) extracted
 IoU: 0.960

Figure 6.4: The Document Explorer view (Macro-Level), demonstrating the side-by-side structural alignment of predicted Agenda Items against the ground truth.

Subject Alignment

Ground Truth Subjects

Alandroal_cm_011_2024-04-24_1
 General Administration, Finance, and Human Resources
 Informação sobre o relatório da ROC em relação à prestação de contas de 2023
 O Sr. Presidente deu conhecimento do relatório da ROC em relação à Prestação de contas de 2023.
 IoU: 0.979

Alandroal_cm_011_2024-04-24_2
 Culture Communication and Public Relations
 Participação do executivo na apresentação do Festival Soil to Soul
 Continuou por informar que no dia 19 o executivo participou na apresentação do Festival Soil to Soul, em Lisboa, que é já a segunda edição e que terá lugar nos dias 25 e 26 de maio, no Castelo do Alan...
 IoU: 0.981

Alandroal_cm_011_2024-04-24_3
 General Administration, Finance, and Human Resources
 Alterações orçamentais permutativas
 Pelo Sr. Presidente foram presentes a reunião as informações da contabilidade que se anexam à presente ata.
 Ponderado e analisado o assunto o Executivo Municipal deliberou por maioria, com os votos a...
 IoU: 0.994

Alandroal_cm_011_2024-04-24_4
 General Administration, Finance, and Human Resources
 Ausência de alteração orçamental permutativa
 IoU: 0.994

Predicted Subjects

Predicted Subject 1
 General Administration, Finance, and Human Resources
 Relatório da ROC sobre a Prestação de Contas de 2023
 O Sr. Presidente deu conhecimento do relatório da ROC em relação à Prestação de contas de 2023.
 ✓ General Administration, Finance, and Human Resources
 ✗ Theme match
 IoU: 0.979

Predicted Subject 2
 Culture
 Participação do executivo na apresentação do Festival Soil to Soul em Lisboa
 Continuou por informar que no dia 19 o executivo participou na apresentação do Festival Soil to Soul, em Lisboa, que é já a segunda edição e que terá lugar nos dias 25 e 26 de maio, no Castelo do Alan...
 ✓ Culture | — Communication and Public Relations
 ✓ Theme match
 IoU: 0.981

Predicted Subject 3
 General Administration, Finance, and Human Resources
 Ratificação das alterações orçamentais permutativas
 Pelo Sr. Presidente foram presentes a reunião as informações da contabilidade que se anexam à presente ata.
 Ponderado e analisado o assunto o Executivo Municipal deliberou por maioria, com os votos a...
 ✓ General Administration, Finance, and Human Resources
 ✓ Theme match
 IoU: 0.994

Figure 6.5: The Document Explorer view (Micro-Level), detailing the extraction of individual subjects, their semantic theme matching, and boundary IoU scores.

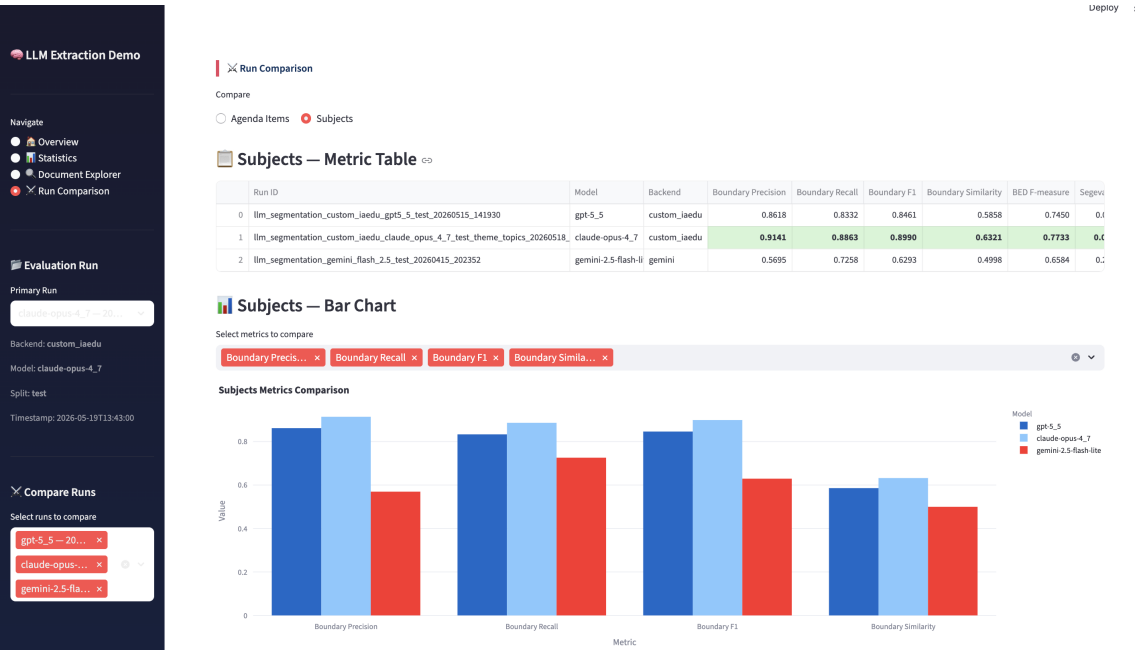


Figure 6.6: The Run Comparison view, providing a visual benchmark of evaluation metrics across different generative model configurations.

tation pipelines were integrated directly into the CitiLink platform⁴ [77], a civic technology portal designed to host and analyze city council meeting minutes.

For this production environment, the platform utilizes the computationally efficient SegNSP model (detailed in Chapter 4) to quickly and reliably parse the incoming meeting minutes. Rather than presenting citizens with dense, multi-page PDF documents, the CitiLink interface leverages the extracted subjects to fundamentally restructure the reading experience.

As demonstrated in the platform’s user interface, documents are dynamically broken down into a navigable repository of discussed subjects (e.g., "Alterações orçamentais permutativas" or "Aprovação de ata"). The platform offers multiple viewing modalities. A grid-based card view (Figure 6.7) allows users to quickly scan the breadth of topics covered in a session, while a chronological list view (Figure 6.8) provides sequential navigation reflecting the actual flow of the meeting.

Users can select any specific subject to open a detailed modal view (Figure 6.9). Here, the isolated transcript corresponding to the model’s boundary predictions is presented alongside an automatically generated summary and a detailed breakdown of the voting distribution (e.g., votes in favor, against, or abstentions).

To ensure the continuous improvement of the underlying AI models, the platform also incorporates user feedback mechanisms (thumbs up/down ratings) on the generated summaries and extracted voting data, establishing a data flywheel for future model fine-tuning.

⁴<https://citilink.inesctec.pt/>

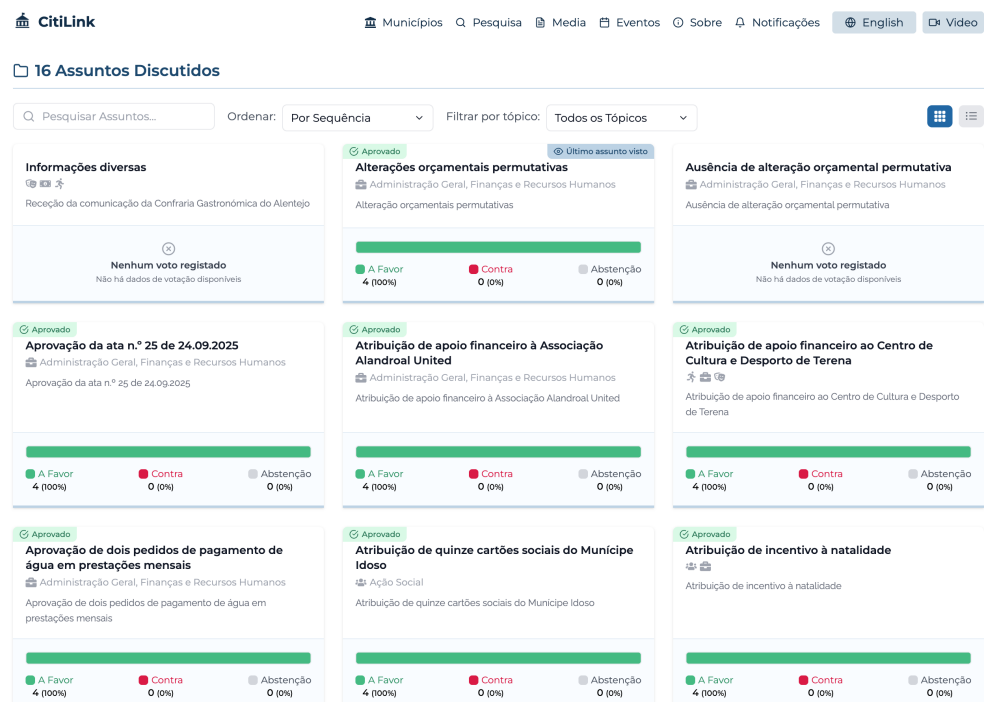


Figure 6.7: The CitiLink platform's grid view, displaying extracted subjects as highly scannable cards equipped with high-level voting distributions.

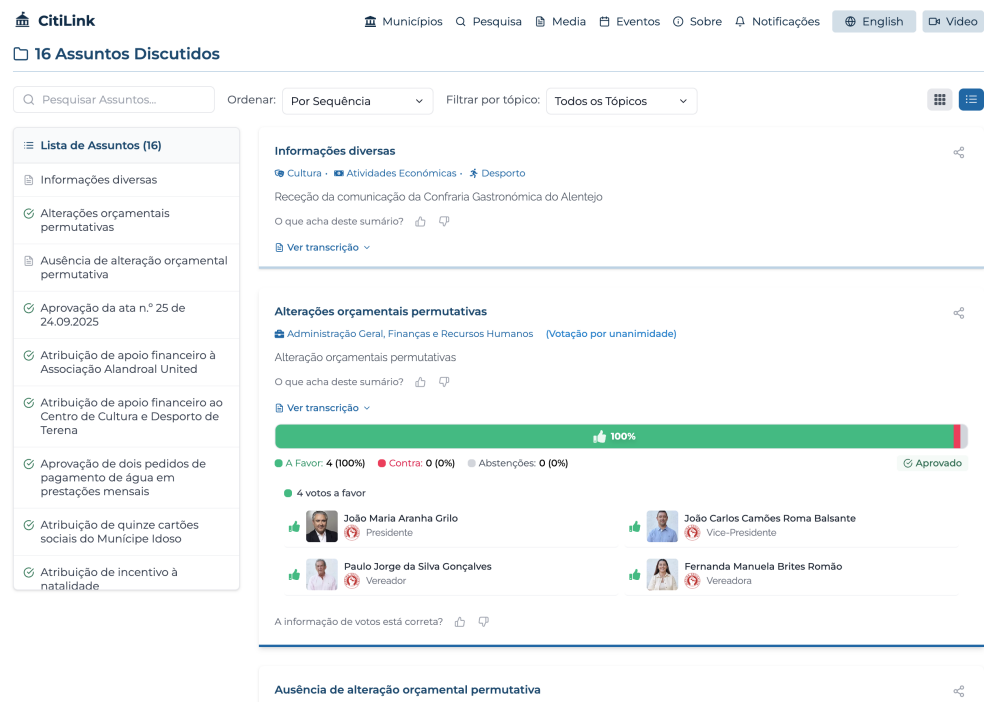


Figure 6.8: The chronological list view in the CitiLink application, enabling seamless navigation between sequentially discussed subjects.



Figure 6.9: Detailed modal view for an individual extracted subject, featuring the generated summary, exact transcript boundaries, and voting breakdown.

6.4 Summary

This chapter demonstrated the transition of the proposed text segmentation models from theoretical artifacts into highly functional, user-centric applications. By developing interactive diagnostic tools for both the SegNSP and HSeg frameworks, we established robust pipelines for the human-in-the-loop qualitative evaluation of NLP models. These visual interfaces proved essential for identifying granular error topologies, such as localized sensitivity to administrative boilerplate and cascading structural hallucinations, which traditional quantitative metrics often mask.

Crucially, the successful integration of the SegNSP model into the CitiLink platform validates the core motivation of this research. By automatically transforming dense municipal minutes into highly structured, navigable, and semantically enriched digital interfaces, the deployment effectively demonstrates that text segmentation is not merely an intrinsic NLP task, but a foundational requirement for modern information retrieval systems. Ultimately, these practical applications prove that applying state-of-the-art segmentation architectures to municipal discourse is both technically viable and highly impactful for enhancing public civic transparency. With the theoretical, empirical, and practical dimensions of this research now fully established, the following chapter concludes the dissertation by summarizing its primary contributions and outlining promising avenues for future work.

Chapter 7

Conclusions and Future Work

This dissertation addressed the challenge of navigating and structuring municipal administrative records, a domain characterized by dense procedural jargon and complex, interleaved thematic discussions. By transitioning the focus of automatic text segmentation from canonical, well-structured encyclopedic datasets to the noisy reality of European Portuguese city council minutes, this work bridges a significant gap in civic technology and applied natural language processing.

Our research progressed through three primary milestones. First, we addressed the fundamental scarcity of domain-specific resources by contributing to the construction of the CitiLink-Minutes dataset. By developing a pipeline to convert flat relational annotations into a standardized, hierarchical schema, we provided the foundational ground truth necessary for training and evaluating administrative natural language processing models.

Second, we introduced SegNSP, demonstrating that linear text segmentation can be effectively modeled as a local discourse continuity problem. By adapting the next sentence prediction objective with a segmentation-aware loss function, SegNSP established a highly competitive linear benchmark on administrative text, proving highly effective at capturing local thematic transitions without the computational overhead of auxiliary topic classification tasks.

Finally, recognizing that linear approaches often falter when confronted with nested administrative boilerplate, we developed HSeg. This recursive, Large Language Model (LLM)-based hierarchical framework successfully mapped macro-level agenda items and micro-level discussion subjects. Crucially, by generalizing the approach to traverse variable-depth n -ary tree structures, we validated HSeg on the massive, open-domain WIKI-727K benchmark.

The results demonstrated that generative models, when constrained by strict sentence-ID grounding and guided by implicit semantic reasoning, establish a fundamentally new overall state-of-the-art. Specifically, HSeg elevated the structural segmentation performance on administrative texts, achieving a micro-level $B-F_1$ of 89.16 with frontier proprietary models. Concurrently, the framework’s integration of open-source models demonstrated the comparative viability of localized, parameter-efficient architectures for institutional proceedings.

Furthermore, on the WIKI-727K benchmark, the framework established new state-of-the-art results across multiple evaluation metrics. Highly optimized lightweight large language mod-

els decisively outperformed the previous best-in-class CRF-based architectures by a substantial absolute margin of +9.06 points in strict F_1 score (80.66), while frontier models simultaneously achieved unprecedented minimums in probabilistic error rates (P_k of 5.30). These findings confirm that large language models significantly outperform traditional sequence-labeling models at capturing deep discourse structures, effectively minimizing both over- and under-segmentation across multiple domains and scales.

Taken together, these contributions demonstrate that the structural challenges posed by European Portuguese municipal minutes, far from being an obstacle to automated analysis, can serve as a productive testbed for advancing both linear and hierarchical segmentation research. By grounding architectural decisions in the specific demands of administrative discourse, this work delivers not only competitive benchmark results, but also a deployable, privacy-compliant pipeline that has already been integrated into a live civic technology platform.

While this work establishes a robust structural foundation for administrative documents, several promising directions remain for extending and applying these contributions further. These span downstream retrieval applications, the development of efficient open-weight alternatives, and the refinement of the hierarchical framework itself, beginning with the integration of segmentation-aware structures into retrieval-augmented generation.

Segmentation-Aware Retrieval-Augmented Generation. A primary direction for subsequent studies is the integration of these topic-aware segmentation models into retrieval-augmented generation pipelines for civic question-answering. As discussed in Sections 1.4 and 2.3, conventional retrieval-augmented systems typically rely on fixed-length or rule-based chunking strategies, which frequently bisect a single subject across multiple chunks, diluting retrieval relevance and increasing the likelihood of hallucinated answers. We hypothesize that replacing these naive splitters with the semantically coherent, structurally-aligned segments produced by SegNSP and HSeg will yield chunks that more faithfully preserve the broader thematic and structural context of a subject, in line with the "Boundary Clarity" and "Chunk Stickiness" principles introduced by Zhao et al. [90]. A natural evaluation framework for this avenue would mirror the downstream summarization experiment of Chapter 4, comparing retrieval precision (e.g., via $DCG@k$) and downstream answer quality when a system is provisioned with fixed-window chunks versus HSeg-derived hierarchical segments. Beyond raw retrieval metrics, this line of work could also explore how the macro-level structure recovered by HSeg might support multi-granularity retrieval, allowing a system to first retrieve a relevant high-level section before drilling down to the specific subject most relevant to a citizen's query.

Fine-Tuning Open-Weight Models for Administrative Segmentation. As discussed in Chapter 5, the evaluation of open-weight and open-source models on CitiLink-Minutes revealed a substantial gap relative to frontier proprietary large language models, particularly for models such as Assistente Multimodal Automático de Linguagem com Inteligência Artificial (AMALIA) (9B)

and Ministral-8B, which struggled to generalize their zero-shot reasoning capabilities to the structural extraction task. This gap, however, was assessed exclusively in a zero-shot setting; the models were never explicitly adapted to the segmentation objective or the procedural vocabulary of Portuguese municipal minutes. A promising direction for future work is therefore the targeted fine-tuning of compact, open-weight models (in the 7B to 12B parameter range) on the structural extraction task defined by HSeg, using the CitiLink-Minutes training split as supervision. Such fine-tuning could draw on parameter-efficient techniques (e.g., LoRA-based adapters) to remain computationally tractable, and could additionally incorporate the segmentation-aware loss formulation introduced for SegNSP in Chapter 4, adapted to the generative setting. Beyond the immediate goal of closing the performance gap with proprietary frontier models, such fine-tuned models would directly address the data privacy and operational cost concerns raised in Section 4.3.5, offering municipalities and civic technology platforms a viable path toward locally-hosted, cost-effective hierarchical segmentation.

Adaptive and Municipality-Aware Calibration. The cross-municipality evaluation in Chapter 4 revealed substantial variance in SegNSP’s performance across different administrative bodies, with $B-F_1$ ranging from 77.68 (Campo Maior) to 24.83 (Fundão). This divergence was largely attributed to municipality-specific formatting conventions, such as highly compressed bullet-point enumerations, which shift the typical distribution of segment lengths. Future work could explore adaptive thresholding mechanisms, where the decision threshold τ is calibrated per municipality, either through a lightweight few-shot calibration step or through a learned threshold predictor conditioned on document-level structural features (e.g., average sentence length, punctuation density). Such mechanisms would help ensure that segmentation systems remain robust as the CitiLink platform onboards new municipalities with previously unseen documentation styles.

Exploiting Existing Sub-Subject Annotations for Deeper Hierarchies. While HSeg was evaluated as a two-level hierarchy on CitiLink-Minutes, this reflects the dataset’s structure at the time the HSeg experiments were conducted, rather than an inherent limitation of the framework itself. As detailed in Chapter 3, the subsequent CitiLink-Minutes v1.1.0 release introduced a `sub_subjects` structure, capturing deeper nested deliberations and their associated sub-voting records within a subset of subjects (Table 3.1). This third structural layer became available only after the HSeg pipeline presented in Chapter 5 had already been designed and evaluated against the dataset’s prior two-level schema, and as such, it was not yet incorporated into the recursive extraction process. A natural avenue for future work is therefore to extend the HSeg evaluation to this newly available third level (agenda items, subjects, and sub-subjects), assessing whether the recursive termination criterion described in Section 5.2, which determines when no further sub-sections exist within a given span, generalizes effectively from the open-domain, variable-depth structures of WIKI-727K to the comparatively sparse and unevenly distributed sub-subject annotations of CitiLink-Minutes (present in only 224 of 2,885 subjects). This would not only provide a more complete evaluation of HSeg’s architectural flexibility, but could also surface whether the

implicit reasoning scaffold explored in Section 5.4 extends naturally to a third level of semantic granularity.

References

- [1] Mistral AI. *Ministral-8B-Instruct-2410*. <https://huggingface.co/mistralai/Ministral-8B-Instruct-2410>. 2024.
- [2] Anthropic. *Claude 4.7 Opus*. <https://www.anthropic.com/news/claude-opus-4-7>. Large Language Model. 2026.
- [3] Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. “SECTOR: A Neural Model for Coherent Topic Segmentation and Classification”. In: *Transactions of the Association for Computational Linguistics* 7 (2019). Ed. by Lillian Lee, Mark Johnson, Brian Roark, and Ani Nenkova, pp. 169–184. DOI: [10.1162/tacl_a_00261](https://doi.org/10.1162/tacl_a_00261). URL: <https://aclanthology.org/Q19-1011/>.
- [4] Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. “Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2721–2727. URL: <https://aclanthology.org/L16-1432/>.
- [5] Dennis Aumiller, Satya Almasian, Sebastian Lackner, and Michael Gertz. “Structural text segmentation of legal documents”. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL ’21. ACM, June 2021, pp. 2–11. DOI: [10.1145/3462757.3466085](https://doi.org/10.1145/3462757.3466085). URL: <http://dx.doi.org/10.1145/3462757.3466085>.
- [6] Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. “Attention-Based Neural Text Segmentation”. In: *Advances in Information Retrieval*. Ed. by Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury. Cham: Springer International Publishing, 2018, pp. 180–193. ISBN: 978-3-319-76941-7.
- [7] Haitao Bai, Pinghui Wang, Ruofei Zhang, and Zhou Su. “SegFormer: A Topic Segmentation Model with Controllable Range of Attention”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.11 (June 2023), pp. 12545–12552. DOI: [10.1609/aaai.v37i11.26477](https://doi.org/10.1609/aaai.v37i11.26477). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/26477>.
- [8] Mostafa Bayomi and Séamus Lawless. “C-HTS: A Concept-based Hierarchical Text Segmentation approach”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and

- Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: <https://aclanthology.org/L18-1241/>.
- [9] Mostafa Bayomi, Killian Levacher, M. Rami Ghorab, and Séamus Lawless. “OntoSeg: A Novel Approach to Text Segmentation Using Ontological Similarity”. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. 2015, pp. 1274–1283. DOI: [10.1109/ICDMW.2015.6](https://doi.org/10.1109/ICDMW.2015.6).
 - [10] Doug Beeferman, Adam Berger, and John Lafferty. “Text Segmentation Using Exponential Models”. In: *Second Conference on Empirical Methods in Natural Language Processing*. 1997. URL: <https://aclanthology.org/W97-0304/>.
 - [11] Iz Beltagy, Matthew E Peters, and Arman Cohan. “Longformer: The long-document transformer”. In: *arXiv preprint arXiv:2004.05150* (2020).
 - [12] Ricardo Campos, Ana Filipa Pacheco, Ana Luísa Fernandes, Inês Cantante, Rute Rebouças, Luís Filipe Cunha, José Isidro, José Evans, Miguel Marques, Rodrigo Batista, Evelin Amorim, Alípio Jorge, Nuno Guimarães, Sérgio Nunes, António Leal, and Purificação Silvano. “CitiLink-Minutes: A Multilayer Annotated Dataset of Municipal Meeting Minutes”. In: *Advances in Information Retrieval*. Ed. by Ricardo Campos, Adam Jatowt, Yanyan Lan, Mohammad Aliannejadi, Christine Bauer, Sean MacAvaney, Avishek Anand, Zhaochun Ren, Suzan Verberne, Nan Bai, and Masoud Mansoury. Cham: Springer Nature Switzerland, 2026, pp. 511–527. ISBN: 978-3-032-21321-1. DOI: [10.1007/978-3-032-21321-1_56](https://doi.org/10.1007/978-3-032-21321-1_56).
 - [13] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain Mccowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. “The AMI meeting corpus: A pre-announcement”. In: July 2005. ISBN: 978-3-540-32549-9. DOI: [10.1007/11677482_3](https://doi.org/10.1007/11677482_3).
 - [14] Branden Chan, Stefan Schweter, and Timo Möller. “German’s Next Language Model”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Doria Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6788–6796. DOI: [10.18653/v1/2020.coling-main.598](https://doi.org/10.18653/v1/2020.coling-main.598). URL: <https://aclanthology.org/2020.coling-main.598/>.
 - [15] Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu, and Dong Yu. “Toward Unifying Text Segmentation and Long Document Summarization”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 106–118. DOI: [10.18653/v1/2022.emnlp-main.8](https://doi.org/10.18653/v1/2022.emnlp-main.8). URL: <https://aclanthology.org/2022.emnlp-main.8/>.
 - [16] Freddy Y. Y. Choi. “Advances in domain independent linear text segmentation”. In: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. 2000. URL: <https://aclanthology.org/A00-2004/>.
 - [17] Michael H. Coen. *When F1 Fails: Granularity-Aware Evaluation for Dialogue Topic Segmentation*. 2025. arXiv: [2512.17083](https://arxiv.org/abs/2512.17083) [cs.CL]. URL: <https://arxiv.org/abs/2512.17083>.
 - [18] Gheorghe Comanici et al. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. 2025. arXiv: [2507.06261](https://arxiv.org/abs/2507.06261) [cs.CL]. URL: <https://arxiv.org/abs/2507.06261>.

- [19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: [10.18653/v1/2020.acl-main.747](https://doi.org/10.18653/v1/2020.acl-main.747). URL: <https://aclanthology.org/2020.acl-main.747/>.
- [20] Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michał Stęchły, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. *mlco2/codecarbon: v2.4.1*. Version v2.4.1. May 2024. DOI: [10.5281/zenodo.11171501](https://doi.org/10.5281/zenodo.11171501). URL: <https://doi.org/10.5281/zenodo.11171501>.
- [21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.
- [22] André V. Duarte, João DS Marques, Miguel Graça, Miguel Freire, Lei Li, and Arlindo L. Oliveira. “LumberChunker: Long-Form Narrative Document Segmentation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 6473–6486. DOI: [10.18653/v1/2024.findings-emnlp.377](https://doi.org/10.18653/v1/2024.findings-emnlp.377). URL: <https://aclanthology.org/2024.findings-emnlp.377/>.
- [23] Jacob Eisenstein. “Hierarchical Text Segmentation from Multi-Scale Lexical Cohesion”. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Ed. by Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 353–361. URL: <https://aclanthology.org/N09-1040/>.
- [24] Tomaz Erjavec, Matyás Kopp, Nikola Ljubesic, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çagri Çöltekin, Danijel Korzinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkađur Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Dargis, Sascha Diwersy, Maria Gavrilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Edward Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fiser. “ParlaMint II: advancing comparable parliamentary corpora across Europe”. In: *Lang. Resour. Evaluation* 59.3 (2025), pp. 2071–2102. DOI: [10.1007/s10579-024-09798-w](https://doi.org/10.1007/s10579-024-09798-w). URL: <https://doi.org/10.1007/s10579-024-09798-w>.
- [25] Daniele Faraglia and Other Contributors. *Faker*. URL: <https://github.com/joke2k/faker>.
- [26] Chris Fournier. “Evaluating Text Segmentation using Boundary Edit Distance”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Hinrich Schuetze, Pascale Fung, and Massimo Poesio. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1702–1712. URL: <https://aclanthology.org/P13-1167/>.

- [27] Steffen Freisinger, Philipp Seeberger, Thomas Ranzenberger, Tobias Bocklet, and Korbinian Riedhammer. “Towards Multi-Level Transcript Segmentation: LoRA Fine-Tuning for Table-of-Contents Generation”. In: *Interspeech 2025*. 2025, pp. 276–280. DOI: [10.21437/Interspeech.2025-2792](https://doi.org/10.21437/Interspeech.2025-2792).
- [28] Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. “Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 11908–11941. DOI: [10.18653/v1/2024.emnlp-main.665](https://doi.org/10.18653/v1/2024.emnlp-main.665). URL: <https://aclanthology.org/2024.emnlp-main.665/>.
- [29] Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. “Discourse Segmentation of Multi-Party Conversation”. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, July 2003, pp. 562–569. DOI: [10.3115/1075096.1075167](https://doi.org/10.3115/1075096.1075167). URL: <https://aclanthology.org/P03-1071/>.
- [30] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: [2312.10997](https://arxiv.org/abs/2312.10997) [cs.CL]. URL: <https://arxiv.org/abs/2312.10997>.
- [31] Gemma Team, Google DeepMind. *Gemma 4: Frontier Multimodal Intelligence on Device*. https://ai.google.dev/gemma/docs/core/model_card_4. Model variant: google/gemma-4-12B-it. 2026.
- [32] Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. “Comparing neural sentence encoders for topic segmentation across domains: not your typical text similarity task”. en. In: *PeerJ Comput Sci* 9 (Nov. 2023), e1593.
- [33] Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. “Recent Trends in Linear Text Segmentation: A Survey”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3084–3095. DOI: [10.18653/v1/2024.findings-emnlp.174](https://doi.org/10.18653/v1/2024.findings-emnlp.174). URL: <https://aclanthology.org/2024.findings-emnlp.174/>.
- [34] Reshmi Ghosh, Harjeet Singh Kajal, Sharanya Kamath, Dhuri Shrivastava, Samyadeep Basu, Hansi Zeng, and Soundararajan Srinivasan. “Topic Segmentation of Semi-structured and Unstructured Conversational Datasets Using Language Models”. In: *Intelligent Systems and Applications*. Ed. by Kohei Arai. Cham: Springer Nature Switzerland, 2024, pp. 91–104. ISBN: 978-3-031-47718-8.
- [35] Dimitrios C. Gklezakos, Timothy Misiak, and Diamond Bishop. *TreeSeg: Hierarchical Topic Segmentation of Large Transcripts*. 2024. arXiv: [2407.12028](https://arxiv.org/abs/2407.12028) [cs.CL]. URL: <https://arxiv.org/abs/2407.12028>.
- [36] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. “Unsupervised Text Segmentation Using Semantic Relatedness Graphs”. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Ed. by Claire Gardent, Raffaella Bernardi, and Ivan Titov. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 125–130. DOI: [10.18653/v1/S16-2016](https://doi.org/10.18653/v1/S16-2016). URL: <https://aclanthology.org/S16-2016/>.

- [37] Google DeepMind. *Gemini 3.1 Flash Lite*. <https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/>. Large Language Model. 2026.
- [38] Nilesh Gupta, Wei-Cheng Chang, Ngot Bui, Cho-Jui Hsieh, and Inderjit S. Dhillon. *LLM-guided Hierarchical Search for End-to-end Reasoning Intensive Retrieval*. 2026. arXiv: 2510.13217 [cs.IR]. URL: <https://arxiv.org/abs/2510.13217>.
- [39] M. A. K. Halliday and R. Hasan. *Cohesion in English*. London: Longman, 1976.
- [40] Marti A. Hearst. “TextTiling: segmenting text into multi-paragraph subtopic passages”. In: *Comput. Linguist.* 23.1 (Mar. 1997), pp. 33–64. ISSN: 0891-2017.
- [41] Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. “MeetingBank: A Benchmark Dataset for Meeting Summarization”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 16409–16423. DOI: 10.18653/v1/2023.acl-long.906. URL: <https://aclanthology.org/2023.acl-long.906/>.
- [42] International Organization for Standardization. *Language resource management — Semantic annotation framework — Part 6: Principles of semantic annotation (SemAF Principles)*. <https://www.iso.org/standard/60581.html>. ISO 24617-6:2016. 2016.
- [43] José Isidro, Filipe Cunha, Purificação Silvano, Alípio Jorge, Nuno Guimarães, Sérgio Nunes, and Ricardo Campos. *SegNSP: Revisiting Next Sentence Prediction for Linear Text Segmentation*. 2026. arXiv: 2601.03474 [cs.CL]. URL: <https://arxiv.org/abs/2601.03474>.
- [44] Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. “Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4859–4870. DOI: 10.18653/v1/2020.acl-main.439. URL: <https://aclanthology.org/2020.acl-main.439/>.
- [45] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. “Dense Passage Retrieval for Open-Domain Question Answering”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. URL: <https://aclanthology.org/2020.emnlp-main.550/>.
- [46] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. “The INCEption Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation”. en. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018). Santa Fe, USA: Association for Computational Linguistics, June 2018, pp. 5–9. URL: <http://tubiblio.ulb.tu-darmstadt.de/106270/>.

- [47] Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. “Text Segmentation as a Supervised Learning Task”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 469–473. DOI: [10.18653/v1/N18-2075](https://doi.org/10.18653/v1/N18-2075). URL: <https://aclanthology.org/N18-2075/>.
- [48] Jeonghwan Lee, Jiyeong Han, Sunghoon Baek, and Min Song. *Topic Segmentation Model Focusing on Local Context*. 2023. arXiv: [2301.01935](https://arxiv.org/abs/2301.01935) [cs.CL]. URL: <https://arxiv.org/abs/2301.01935>.
- [49] Seungmin Lee, Yongsang Yoo, Minhwa Jung, and Min Song. “Def-DTS: Deductive Reasoning for Open-domain Dialogue Topic Segmentation”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 20736–20753. ISBN: 979-8-89176-256-5. DOI: [10.18653/v1/2025.findings-acl.1066](https://doi.org/10.18653/v1/2025.findings-acl.1066). URL: <https://aclanthology.org/2025.findings-acl.1066/>.
- [50] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: [10.18653/v1/2020.acl-main.703](https://doi.org/10.18653/v1/2020.acl-main.703). URL: <https://aclanthology.org/2020.acl-main.703/>.
- [51] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. “Retrieval-augmented generation for knowledge-intensive NLP tasks”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS ’20*. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.
- [52] Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. “Molweni: A Challenge Multiparty Dialogues-based Machine Reading Comprehension Dataset with Discourse Structure”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2642–2652. DOI: [10.18653/v1/2020.coling-main.238](https://doi.org/10.18653/v1/2020.coling-main.238). URL: <https://aclanthology.org/2020.coling-main.238/>.
- [53] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. “Focal Loss for Dense Object Detection”. In: vol. 42. 2. 2020, pp. 318–327. DOI: [10.1109/TPAMI.2018.2858826](https://doi.org/10.1109/TPAMI.2018.2858826).
- [54] Yang Liu, Chenguang Zhu, and Michael Zeng. “End-to-End Segmentation-based News Summarization”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 544–554. DOI: [10.18653/v1/2022.findings-acl.46](https://doi.org/10.18653/v1/2022.findings-acl.46). URL: <https://aclanthology.org/2022.findings-acl.46/>.

- [55] Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray Buntine. “Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3334–3340. DOI: [10.18653/v1/2021.findings-emnlp.283](https://doi.org/10.18653/v1/2021.findings-emnlp.283). URL: <https://aclanthology.org/2021.findings-emnlp.283/>.
- [56] Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonalo Simoes. “Text Segmentation by Cross Segment Attention”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 4707–4716. DOI: [10.18653/v1/2020.emnlp-main.380](https://doi.org/10.18653/v1/2020.emnlp-main.380). URL: <https://aclanthology.org/2020.emnlp-main.380/>.
- [57] Itzik Malkiel, Uri Alon, Yakir Yehuda, Shahar Keren, Oren Barkan, Royi Ronen, and Noam Koenigstein. “SEGLLM: Topic-Oriented Call Segmentation Via LLM-Based Conversation Synthesis”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 11361–11365. DOI: [10.1109/ICASSP48485.2024.10446156](https://doi.org/10.1109/ICASSP48485.2024.10446156).
- [58] Miguel Marques, Ana Lusa Fernandes, Ana Filipa Pacheco, Rute Rebouas, Ins Cantante, Jos Isidro, Lus Filipe Cunha, Alpio Jorge, Nuno Guimares, Srgio Nunes, Antnio Leal, Purificao Silvano, and Ricardo Campos. “CitiLink-Summ: A Dataset of Discussion Subjects Summaries in European Portuguese Municipal Meeting Minutes”. In: *Proceedings of the ACM Web Conference 2026*. WWW ’26. United Arab Emirates: Association for Computing Machinery, 2026, pp. 8705–8708. ISBN: 9798400723070. DOI: [10.1145/3774904.3792945](https://doi.org/10.1145/3774904.3792945). URL: <https://doi.org/10.1145/3774904.3792945>.
- [59] Pedro Henrique Martins, Patrick Fernandes, Joo Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, Jos Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, Jos G. C. de Souza, Alexandra Birch, and Andr F. T. Martins. *EuroLLM: Multilingual Language Models for Europe*. 2024. arXiv: [2409.16235](https://arxiv.org/abs/2409.16235) [cs.CL]. URL: <https://arxiv.org/abs/2409.16235>.
- [60] Nicolas Merz, Sven Regel, and Jirka Lewandowski. “The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis”. In: *Research & Politics* 3.2 (2016), p. 2053168016643346. DOI: [10.1177/2053168016643346](https://doi.org/10.1177/2053168016643346). eprint: <https://doi.org/10.1177/2053168016643346>. URL: <https://doi.org/10.1177/2053168016643346>.
- [61] Jane Morris and Graeme Hirst. “Lexical Cohesion Computed by Thesaural relations as an indicator of the structure of text”. In: *Computational Linguistics* 17.1 (1991), pp. 21–48. URL: <https://aclanthology.org/J91-1002/>.
- [62] Inderjeet Nair, Aparna Garimella, Balaji Vasan Srinivasan, Natwar Modani, Niyati Chhaya, Srikrishna Karanam, and Sumit Shekhar. “A Neural CRF-based Hierarchical Approach for Linear Text Segmentation”. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 883–893. DOI: [10.18653/v1/2023.findings-eacl.65](https://doi.org/10.18653/v1/2023.findings-eacl.65). URL: <https://aclanthology.org/2023.findings-eacl.65/>.

- [63] Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. “ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, June 2022, pp. 3174–3182. URL: <https://aclanthology.org/2022.lrec-1.340/>.
- [64] Hai Toan Nguyen, Tien Dat Nguyen, and Viet Ha Nguyen. *Enhancing Retrieval Augmented Generation with Hierarchical Text Segmentation Chunking*. 2025. arXiv: 2507.09935 [cs.CL]. URL: <https://arxiv.org/abs/2507.09935>.
- [65] Hai-Toan Nguyen, Tien-Dat Nguyen, and Viet-Ha Nguyen. “Enhancing Retrieval Augmented Generation with Hierarchical Text Segmentation Chunking”. In: *Information and Communication Technology*. Ed. by Wray Buntine, Morten Fjeld, Truyen Tran, Minh-Triet Tran, Binh Huynh Thi Thanh, and Takumi Miyoshi. Singapore: Springer Nature Singapore, 2025, pp. 209–220. ISBN: 978-981-96-4288-5.
- [66] Tongke Ni, Yang Fan, Junru Zhou, Xiangping Wu, and Qingcai Chen. “CrossFormer: Cross-Segment Semantic Fusion for Document Segmentation”. In: (Apr. 2025). URL: <http://arxiv.org/abs/2503.23671>.
- [67] Maciej Ogrodniczuk. “Polish parliamentary corpus”. In: *Proceedings of the LREC 2018 workshop ParlaCLARIN: creating and using parliamentary corpora*. 2018, pp. 15–19.
- [68] Hyo-Jung Oh, Sung Hyon Myaeng, and Myung-Gil Jang. “Semantic passage segmentation based on sentence topics for question answering”. In: *Information Sciences* 177.18 (2007), pp. 3696–3717. ISSN: 0020-0255. DOI: <https://doi.org/10.1016/j.ins.2007.02.038>. URL: <https://www.sciencedirect.com/science/article/pii/S0020025507001302>.
- [69] OpenAI. *GPT-5.5*. <https://openai.com/index/gpt-5-5-system-card/>. Large Language Model. 2026.
- [70] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. *Regularizing Neural Networks by Penalizing Confident Output Distributions*. 2017. arXiv: 1701.06548 [cs.NE]. URL: <https://arxiv.org/abs/1701.06548>.
- [71] Lev Pevzner and Marti A. Hearst. “A Critique and Improvement of an Evaluation Metric for Text Segmentation”. In: *Computational Linguistics* 28.1 (2002), pp. 19–36. DOI: [10.1162/089120102317341756](https://doi.org/10.1162/089120102317341756). URL: <https://aclanthology.org/J02-1002/>.
- [72] Matthew Purver. “Topic segmentation”. In: *Spoken language understanding: systems for extracting semantic information from speech* (2011), pp. 291–317.
- [73] Renyi Qu, Ruixuan Tu, and Forrest Sheng Bao. “Is Semantic Chunking Worth the Computational Cost?” In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 2155–2177. ISBN: 979-8-89176-195-7. DOI: [10.18653/v1/2025.findings-naacl.114](https://doi.org/10.18653/v1/2025.findings-naacl.114). URL: <https://aclanthology.org/2025.findings-naacl.114/>.

- [74] Fabian Retkowski and Alexander Waibel. “From Text Segmentation to Smart Chaptering: A Novel Benchmark for Structuring Video Transcriptions”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 406–419. DOI: [10.18653/v1/2024.eacl-long.25](https://doi.org/10.18653/v1/2024.eacl-long.25). URL: <https://aclanthology.org/2024.eacl-long.25/>.
- [75] Martin Riedl and Chris Biemann. “TopicTiling: A Text Segmentation Algorithm based on LDA”. In: *Proceedings of ACL 2012 Student Research Workshop*. Ed. by Jackie C. K. Cheung, Jun Hatori, Carlos Henriquez, and Ann Irvine. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 37–42. URL: <https://aclanthology.org/W12-3307/>.
- [76] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. “The ICSI Meeting Recorder Dialog Act (MRDA) Corpus”. In: *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. Cambridge, Massachusetts, USA: Association for Computational Linguistics, Apr. 2004, pp. 97–100. URL: <https://aclanthology.org/W04-2319/>.
- [77] Rodrigo Silva, José Evans, José Isidro, Miguel Marques, Afonso Fonseca, Ricardo Morais, João Canavilhas, Arian Pasquali, Purificação Silvano, Alípio Jorge, Nuno Guimarães, Sérgio Nunes, and Ricardo Campos. “CitiLink: Enhancing Municipal Transparency and Citizen Engagement Through Searchable Meeting Minutes”. In: *Advances in Information Retrieval*. Ed. by Ricardo Campos, Adam Jatowt, Yanyan Lan, Mohammad Aliannejadi, Christine Bauer, Sean MacAvaney, Avishek Anand, Zhaochun Ren, Suzan Verberne, Nan Bai, and Masoud Mansoury. Cham: Springer Nature Switzerland, 2026, pp. 193–199. ISBN: 978-3-032-21321-1. DOI: [10.1007/978-3-032-21321-1_28](https://doi.org/10.1007/978-3-032-21321-1_28).
- [78] Afonso Simplício, Gonçalo Vinagre, Miguel Moura Ramos, Diogo Tavares, Rafael Ferreira, Giuseppe Attanasio, Duarte M. Alves, Inês Calvo, Inês Vieira, Rui Guerra, James Furtado, Beatriz Canaverde, Iago Paulo, Vasco Ramos, Diogo Glória-Silva, Miguel Faria, Marcos Treviso, Daniel Gomes, Pedro Gomes, David Semedo, André Martins, and João Magalhães. “AMALIA: A Fully Open Large Language Model for European Portuguese”. In: *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*. Ed. by Marlo Souza, Iria de-Dios-Flores, Diana Santos, Larissa Freitas, Jackson Wilke da Cruz Souza, and Eugénio Ribeiro. Salvador, Brazil: Association for Computational Linguistics, Apr. 2026, pp. 380–391. ISBN: 979-8-89176-387-6. URL: <https://aclanthology.org/2026.propor-1.38/>.
- [79] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. “BERTimbau: pretrained BERT models for Brazilian Portuguese”. In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. 2020.
- [80] Jörg Tiedemann and Jori Mur. “Simple is Best: Experiments with Different Document Segmentation Strategies for Passage Retrieval”. In: *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*. Ed. by Mark A. Greenwood. Manchester, UK: Coling 2008 Organizing Committee, Aug. 2008, pp. 17–25. URL: <https://aclanthology.org/W08-1803/>.
- [81] Paveen Virameteekul. “Paragraph-level attention based deep model for chapter segmentation”. In: *PeerJ Computer Science* 8 (2022). ISSN: 23765992. DOI: [10.7717/PEERJ-CS.1003](https://doi.org/10.7717/PEERJ-CS.1003). URL: <https://peerj.com/articles/cs-1003/>.

- [82] Zhitong Wang, Cheng Gao, Chaojun Xiao, Yufei Huang, Shuzheng Si, Kangyang Luo, Yuzhuo Bai, Wenhao Li, Tangjian Duan, Chuancheng Lv, Guoshan Lu, Gang Chen, Fan-chao Qi, and Maosong Sun. “Document Segmentation Matters for Retrieval-Augmented Generation”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 8063–8075. ISBN: 979-8-89176-256-5. DOI: [10.18653/v1/2025.findings-acl.422](https://doi.org/10.18653/v1/2025.findings-acl.422). URL: <https://aclanthology.org/2025.findings-acl.422/>.
- [83] Sabine Wehnert, Harikrishnan Changaramkulath, and Ernesto William De Luca. *HiPS: Hierarchical PDF Segmentation of Textbooks*. 2025. arXiv: 2509.00909 [cs.IR]. URL: <https://arxiv.org/abs/2509.00909>.
- [84] Claes Wohlin. “Guidelines for snowballing in systematic literature studies and a replication in software engineering”. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. EASE ’14. London, England, United Kingdom: Association for Computing Machinery, 2014. ISBN: 9781450324762. DOI: [10.1145/2601248.2601268](https://doi.org/10.1145/2601248.2601268). URL: <https://doi.org/10.1145/2601248.2601268>.
- [85] Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. “TIAGE: A Benchmark for Topic-Shift Aware Dialog Modeling”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1684–1690. DOI: [10.18653/v1/2021.findings-emnlp.145](https://doi.org/10.18653/v1/2021.findings-emnlp.145). URL: <https://aclanthology.org/2021.findings-emnlp.145/>.
- [86] Linzi Xing and Giuseppe Carenini. “Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring”. In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitrallekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li. Singapore and Online: Association for Computational Linguistics, July 2021, pp. 167–177. DOI: [10.18653/v1/2021.sigdial-1.18](https://doi.org/10.18653/v1/2021.sigdial-1.18). URL: <https://aclanthology.org/2021.sigdial-1.18/>.
- [87] Jiahui Xu, Feng Jiang, Anningzhe Gao, Luis Fernando D’Haro, and Haizhou Li. “Unsupervised Mutual Learning of Discourse Parsing and Topic Segmentation in Dialogue”. In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 4667–4681. DOI: [10.1109/TASLPRO.2025.3623884](https://doi.org/10.1109/TASLPRO.2025.3623884).
- [88] Yi Xu, Hai Zhao, and Zhuosheng Zhang. “Topic-Aware Multi-turn Dialogue Modeling”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.16 (May 2021), pp. 14176–14184. DOI: [10.1609/aaai.v35i16.17668](https://doi.org/10.1609/aaai.v35i16.17668). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/17668>.
- [89] An Yang et al. “Qwen2 Technical Report”. In: *arXiv preprint arXiv:2407.10671* (2024).
- [90] Jihao Zhao, Zhiyuan Ji, Zhaoxin Fan, Hanyu Wang, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. “MoC: Mixtures of Text Chunking Learners for Retrieval-Augmented Generation System”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 5172–5189. ISBN: 979-8-89176-251-0. DOI: [10.18653/v1/](https://doi.org/10.18653/v1/)

2025.acl-long.258. URL: <https://aclanthology.org/2025.acl-long.258/>.

Appendix A

Supplementary Evaluation Results

This appendix provides the supplementary cross-domain and cross-lingual segmentation results for SegNSP on the German WikiSection subsets (`de_city` and `de_disease`). These results were omitted from the main text in Chapter 4 to improve table readability and focus the primary discussion on the English and Portuguese datasets.

The evaluation reported here follows the exact same standard train/test splits, baseline configurations, and sequence-labeling metrics detailed in the core experimental setup.

Table A.1: Cross-domain segmentation results for the German WikiSection subsets. Higher B - F_1/B and lower P_k/WD indicate better performance. Values are macro-averaged; best in bold.

Model	de_city				de_disease			
	B - F_1 $\uparrow$	B $\uparrow$	P_k $\downarrow$	WD $\downarrow$	B - F_1 $\uparrow$	B $\uparrow$	P_k $\downarrow$	WD $\downarrow$
TextTiling	16.45	9.41	33.03	33.13	13.41	7.04	36.78	36.90
Att+CNN	22.88	14.25	30.60	32.00	11.32	6.65	38.86	39.19
TopSeg	27.90	16.57	18.53	22.42	19.85	11.16	29.65	32.43
LumberChunker	–	–	–	–	–	–	–	–
STP + TC + NSP	–	–	–	–	–	–	–	–
SegNSP	60.09	46.69	17.02	21.43	52.00	42.20	22.13	27.10

Appendix B

HSeg Prompt Templates

This section details the exact prompt templates and few-shot examples used to constrain the generative LLMs in the HSeg pipeline.

```

STAGE1_PROMPT_TEMPLATE = """You are an expert at analyzing Portuguese municipal council
↪ meeting minutes. Your task is to identify agenda items in the document.

INSTRUCTIONS:
1. Agenda items are the MAJOR top-level discussion topics, usually numbered ("1.", "2.",
↪ "3.", etc.)
2. Sub-items (e.g., "2.1.", "2.2.", "3.a") are part of their parent item -- do NOT split
↪ them into separate entries
3. Each agenda item spans all its sub-items and associated discussion
4. Look for patterns like:
    - Numbered titles: "1. PEDIDO DE...", "2. APROVAÇÃO...", etc.
    - Formal transitions between major topics
5. A typical Portuguese council meeting has between 3 and 20 major agenda items -- if you
↪ find significantly more, you are likely splitting sub-items incorrectly

INPUT FORMAT:
Below is the full meeting minute document, with each sentence assigned an ID (S1, S2, S3,
↪ ...).

{sentence_mapping}

OUTPUT FORMAT:
For each agenda item you identify, provide:
1. item_title: The title or description of the agenda item (the top-level number and
↪ title)
2. start_id: The sentence ID where the item begins (e.g., S5)
3. end_id: The sentence ID where the item ends (e.g., S12)

Return your answer as a JSON array:
[
    {"item_title": "Title of item 1", "start_id": "S1", "end_id": "S10"}},
    ...
]

IMPORTANT:
- Items must NOT overlap (end_id of item N < start_id of item N+1)
- Items should cover the entire document (no gaps)
- Only use sentence IDs that exist in the input (S1 to S{max_id})
- Sub-items (2.1, 2.2, etc.) belong inside their parent item -- do NOT list them
↪ separately
- Group all sub-item discussion under the parent agenda item

Now, identify the agenda items:"""

```

Listing 4: Stage 1 prompt template for extracting macro-level agenda items in the CitiLink-Minutes dataset.

```

STAGE2_PROMPT_TEMPLATE = """You are an expert at analyzing Portuguese municipal council
↪ meeting minutes. Your task is to identify discussion subjects within a specific agenda
↪ item.

CONTEXT:
Agenda Item: {item_title}

INSTRUCTIONS:
1. A "subject" is a cohesive subtopic of discussion within the agenda item
2. Each subject typically includes:
  - Introduction/presentation of the matter (e.g., "Pelo Senhor Presidente foi
  ↪ presente...")
  - Discussion (if any)
  - Decision/deliberation (e.g., "...deliberou por unanimidade aprovar...")
3. Multiple subjects can exist within one agenda item -- a typical agenda item has 1 to 5
  ↪ subjects
4. Look for transitions that signal a new subject:
  - "Pelo Senhor Presidente foi presente..." / "De seguida..."
  - New proposals, requests, or reports being introduced
  - Different entities, contracts, or matters being discussed
5. Each distinct matter discussed (different request, different approval, different
  ↪ report) is a separate subject

IMPORTANT:
{municipality_hint}

INPUT FORMAT:
Below are the sentences from this agenda item, with IDs:

{sentence_mapping}

OUTPUT FORMAT:
For each subject you identify, provide:{output_format_instructions}
{start_id_instruction}
{end_id_instruction}

Return your answer as a JSON array:
[
  {{
    {example_json_fields}
  }},
  ...
]
{topic_categories_section}

IMPORTANT CONSTRAINTS:
- Subjects must NOT overlap
- Subjects should cover the entire agenda item (minimize gaps)
- Only use sentence IDs from the provided list{topic_constraint}
- If the agenda item has only one subject, return an array with one element
- Each distinct matter (pedido, aprovação, proposta, etc.) should be its own subject

Now, identify the subjects:"""

```

Listing 5: Stage 2 prompt template for extracting micro-level subjects within an agenda item, in the CitiLink-Minutes dataset.

```
STAGE1_ROOT_PROMPT_TEMPLATE = """\
You are an expert at analyzing Wikipedia articles. Your task is to identify the \
top-level sections in the article.

INSTRUCTIONS:
1. Top-level sections are the MAJOR divisions of the article \
(e.g. "History", "Overview", "Background", "Usage", "Architecture").
2. Section headings appear in the text as standalone lines (a short phrase \
on its own line, not ending with a period and not part of a regular sentence). \
Use them as your primary signal for section boundaries.
3. Sub-sections (e.g. "Early history", "Modern usage") belong INSIDE their \
parent section -- do NOT list them as separate top-level items.
4. A typical Wikipedia article has between 3 and 15 major sections.
5. Each section spans all its sub-sections and associated text.

INPUT FORMAT:
Below is the Wikipedia article body with each sentence assigned an ID \
(S1, S2, S3, ...). Standalone heading lines are included as their own sentences:

{sentence_mapping}

OUTPUT FORMAT:
Return a JSON array where each object represents a top-level section:
[
  {"section_title": "Exact heading text", "start_id": "S1", "end_id": "S10"},
  ...
]

CONSTRAINTS:
- Use the heading line's sentence ID as ``start_id`` for that section.
- Sections must NOT overlap.
- Only use sentence IDs that exist in the input (S1 to S{max_id}).
- Sub-sections belong inside their parent -- do NOT list them separately.

Now, identify the top-level sections:\
"""
```

Listing 6: Stage 1 prompt template utilized for hierarchical top-level extraction in Wiki-50 and WIKI-727k documents.