Beyond OCR: Structural Segmentation and Speaker Attribution in Historical Italian Parliamentary Debates

Claudia Corbetta, Samuele Mazzei, Alessio Palmero Aprosio
University of Trento, Corso Bettini 84, Rovereto (Italy)
claudia.corbetta@unitn.it, samuele.mazzei@studenti.unitn.it, a.palmeroaprosio@unitn.it

Abstract

Historical parliamentary debates are essential for longitudinal political and linguistic research, yet much early material remains available only as scanned images. In the Italian context, proceedings from 1848–1996 lack large-scale, structurally annotated, machine-readable representations. This paper addresses the challenge of transforming historical Italian parliamentary debates into structured corpora by moving beyond plain Optical Character Recognition (OCR) toward functional block segmentation and speaker attribution. We present detailed annotation guidelines and a manually annotated dataset of 300 randomly sampled pages. Two approaches are compared: (i) direct multimodal Large Language Model (LLM) annotation and (ii) a modular pipeline combining OCR with LLM-based structural reconstruction under zero-shot and few-shot prompting. Evaluation on a held-out test set shows that separating transcription from structural reasoning improves performance, with few-shot prompting yielding the most reliable results. The study demonstrates the feasibility of integrating LLM-based reasoning into historical parliamentary digitisation workflows.

Keywords: Historical Parliamentary Corpora, Optical Character Recognition (OCR), Speaker Attribution

1. Introduction

Parliamentary debates represent the most comprehensive and continuous record of a nation's political, social, and linguistic evolution, capturing the direct confrontation between ideologies and the formal legislative process as it unfolds.

The study of parliamentary proceedings has been revolutionized by the emergence of large-scale computational text analysis. Systematic infrastructures like the ParlaMint project (Erjavec et al., 2023, 2024) have paved the way for comparative European legislative research by providing harmonized, multilingual datasets. However, there is a significant disparity in the availability of these resources.

Most existing studies rely on recent transcriptions, as modern data is collected digitally and already annotated semantically, making it easily accessible for researchers. In contrast, research on older historical data remains largely "vertical", limited to small time intervals or specific subsets of debates, primarily due to the lack of structured, machine-readable data.

In the Italian context, although recent efforts have produced machine-actionable corpora for contemporary parliamentary data, most historical proceedings from the Kingdom and early Republic remain available only as scanned images. Exploratory initiatives such as IPSA (Frasnelli and Palmero Aprosio, 2024) and more recent structured resources like ItaParlCorpus (Cova, 2025) address parts of this landscape, yet large-scale, structurally annotated representations of long-span historical debates are still missing.

Although established OCR and layout analysis pipelines exist (Breuel, 2008; Wick et al., 2018), they do not integrate higher-level reasoning mechanisms capable of jointly addressing block segmentation and speaker attribution in complex historical parliamentary layouts.

Accurate block segmentation and speaker attribution are essential for downstream analyses of political discourse and representation. Longitudinal studies require distinguishing between communicative units, separating oral interventions from legislative articles, procedural notes, and other structural elements.

This paper addresses the challenges posed by Italian historical parliamentary documents through a multi-step framework centered on the following research question: which computational approach most effectively integrates accurate OCR with reliable structural reconstruction of document layout and speaker segmentation in long-span historical data? To answer this question, we first introduce comprehensive guidelines for the annotation of textual blocks in Italian parliamentary debates, designed to support downstream tasks such as discourse analysis and speaker tracking. These guidelines make explicit the structural complexity and information density of the documents, highlighting how multiple textual layers (discursive, procedural, and editorial) coexist within the same pages. We then present a dataset of 300 pages, randomly sampled from the data and manually annotated accord-

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 65–76

ing to this framework. Building on this resource, we survey the main methodological approaches to the task, encompassing both optical character recognition and block/speaker identification strategies. Finally, we conduct a systematic comparative evaluation, measuring the extent to which different systems succeed not only in accurately transcribing the textual content, but also in reconstructing the structural organization of the documents. This evaluation ultimately allows us to identify the approach that achieves the best balance between textual fidelity and structural comprehension across nearly 150 years of Italian parliamentary proceedings.

2. Related Work

Parliamentary transcripts constitute a fundamental resource for longitudinal research in political science and linguistics, enabling the analysis of ideological trends, agenda-setting, political representation, and the evolution of political discourse over time, as illustrated, for example, by studies focusing on the Italian parliamentary context (Curini et al., 2024; Cominetti et al., 2022). The large-scale digitisation of parliamentary archives across Europe has led to the creation of computational corpora such as Hansard (Wattam et al., 2014; Nanni et al., 2019; Coole et al., 2020), GERPARCOR (Abrami et al., 2022, 2024), and GePaDe (Rehbein et al., 2024), among others, which integrate OCR, structural reconstruction, metadata enrichment, and linguistic annotation to support systematic analysis of legislative debates. Similarly, infrastructures such as Parla-CLARIN[1] and workflows like OCR4all[2] demonstrate the feasibility of transforming historical parliamentary scans into structured, machine-readable corpora through iterative OCR and annotation pipelines (Kavčič et al., 2024). These initiatives highlight the importance of integrating OCR with downstream linguistic processing and structural annotation to enable large-scale computational analysis.

However, the **digitisation of historical parliamentary** records presents significant technical challenges due to the limitations of OCR when applied to degraded materials, non-standard typography, and complex layouts typical of historical printings, which increase recognition errors and complicate text reconstruction (Reul et al., 2019; Greif et al., 2025). In particular, Document Layout Analysis (DLA) remains a critical bottleneck, as errors in segmenting multi-column formats, speaker markers, and procedural elements can disrupt reading order and affect downstream tasks such as speaker attribution and discourse segmentation. To mitigate these issues, recent workflows combine layout segmentation, deep learning–based OCR models, and iterative training with manual correction to improve accuracy and robustness.

Within this context, **annotation workflows** are essential for transforming OCR-derived text into reliable research corpora, as OCR output often contains structural inconsistencies and recognition errors that must be resolved during annotation. For instance, the IsraParlTweet corpus required iterative preprocessing combining rule-based extraction and human validation to accurately identify speakers and segment debates (Mor-Lan et al., 2024), highlighting that annotation is an integral part of the digitisation process rather than a separate downstream task. Clear annotation guidelines and structured schemes are crucial to ensure consistency and resolve ambiguities caused by OCR noise (Reinig et al., 2024), while multilayer workflows incorporating double annotation and expert validation, such as those used in the CitiLink-Minutes dataset, further improve annotation reliability and correct digitisation-induced inconsistencies (Guimarães et al., 2025).

3. Annotation Guidelines

As stated in Section 2, the annotation process represents a critical step in transforming OCR-derived parliamentary transcripts into reliable research corpora, as raw OCR output often contains structural ambiguities, segmentation errors, and incomplete speaker attribution.

The annotation of historical parliamentary debates was therefore carried out following a structured set of guidelines[3] designed to segment transcripts into coherent functional units and assign consistent metadata to each segment. By systematically identifying and labeling functional textual blocks, annotation enables the reconstruction of the logical and communicative structure of parliamentary proceedings, ensuring that speeches, legislative content, and procedural elements can be accurately distinguished and analyzed.

While these guidelines provide a robust framework for the Italian parliamentary context, they have been specifically designed to accommodate the idiosyncratic layout conventions, document structure, and procedural norms of the Italian Parliament. Consequently, although the underlying principles of functional segmentation and metadata annotation are broadly applicable, the schema may require adaptation or extension to address the diplomatic, legal, and linguistic conventions of parliamentary debates in other national contexts.

[1] https://github.com/clarin-eric/parla-clarin

[2] https://github.com/ocr4all

[3] All the material related to this article is available on the main Github of the IPSA project: https://github.com/dhfbk/ipsa

3.1. Annotation workflow

Annotation was performed at the level of textual blocks, defined as continuous stretches of text with a single communicative function, such as a speech turn, legislative article, procedural description, or quoted document.

Each block was assigned a unique instance identifier within the page and annotated with a type label, speaker information, and optional notes to clarify ambiguous cases. Specifically, for each instance, a structured set of annotation fields was defined, including:

- **Page ID**: the corresponding PDF page number;
- **Instance/Round ID**: a progressive integer starting at 1 for each new page;
- **Type Label**: a functional category identifying the communicative nature of the block (refer to Table 1);
- **Speaker**: the individual or institutional role responsible for the speech or text (refer to 3.2);
- **Notes**: optional clarifications, particularly to specify the nature of blocks labeled as "other".

Moreover, the annotation scheme relied on a predefined taxonomy of block types (i.e., type label), including speech, article, text, description, title, and other, to capture the structural and functional diversity of parliamentary records. Table 1 provides a detailed description of the communicative function associated with each label, clarifying the criteria used to distinguish between oral interventions, legislative content, procedural annotations, and other structural elements of the parliamentary proceedings.

Segmentation of block type is based on functional and graphical criteria. A new unit is annotated whenever the text changes its communicative function and this shift is clearly signaled by layout features such as line breaks, indentation, or other formatting cues. Typical cases include the beginning of a new speaker's turn, the introduction of a legislative article, the presence of a procedural note, the appearance of a title, or the insertion of attachments or tables. Consecutive non-speech elements are annotated as separate units whenever they represent distinct functional blocks.[4] However, when multiple components form a single integrated unit, they are annotated as one segment. The proposed annotation schema is partially aligned with existing standards for parliamentary corpora, such as the ParlaMint schema.[5] In particular, core elements such as speaker attribution and speech segmentation can be directly mapped to ParlaMint components (e.g., <u> elements with speaker metadata). However, our annotation framework introduces a finer-grained distinction of functional block types (e.g., description, speechnotext, presidencydeclaration). These categories are motivated by the need to capture the structural and procedural complexity of historical parliamentary documents, where non-speech elements (e.g., procedural notes, legislative articles, titles, and editorial insertions) play a central role in structuring the document. More generally, this design choice reflects a different modelling perspective: while standard parliamentary encoding schemes such as ParlaMint are primarily designed for TEI-compliant textual representation, our framework aims to support structuring and linking of information in a Linked Open Data (LOD) setting. As a result, certain distinctions are made explicit at the annotation level in order to facilitate downstream semantic integration.

3.2. Speaker Attribution

Given the parliamentary nature of the texts, accurate identification of the speaker is a fundamental requirement for enabling reliable analysis of political discourse, speaker behavior, and institutional dynamics. For this reason, particular attention was devoted to the annotation of speaker attribution.

Speaker attribution was designed to make speaker identification as explicit and interpretable as possible. The annotation captures speaker mentions at the textual level, without enforcing cross-document identity resolution, which is deferred to a later linking stage (e.g., via LOD). To this end, a set of labels (explicit, inferred, hidden, unknown) was introduced to represent different degrees of speaker identifiability. More specifically, the identification of the **speaker**, recorded in the *speaker* column, includes both the personal name and any associated institutional role when explicitly provided in the source text:

"Baccarini, ministro dei lavori pubblici";

"'Rosadi".

In cases of generic or collective interjections where no individual speaker can be identified, the prefix `[unknown]` is used:

"`[unknown]` Una voce".

When the speaker is not explicitly named in the current segment but can be reliably inferred from

[4] Similarly, consecutive elements classified as "other" (e.g., attachments, tables, or related materials) are annotated separately if they are visually and structurally distinct.

[5] See https://clarin-eric.github.io/ParlaMint/.

the context (for instance, as a continuation from a previous speech), the prefix `[inferred]` is applied:

"`[inferred]` Presidente".

Conversely, if the speaker cannot be identified, as in cases where the speech is ongoing and the speaker was mentioned on a previous page, the label `[hidden]` is assigned:

"`[hidden]`".

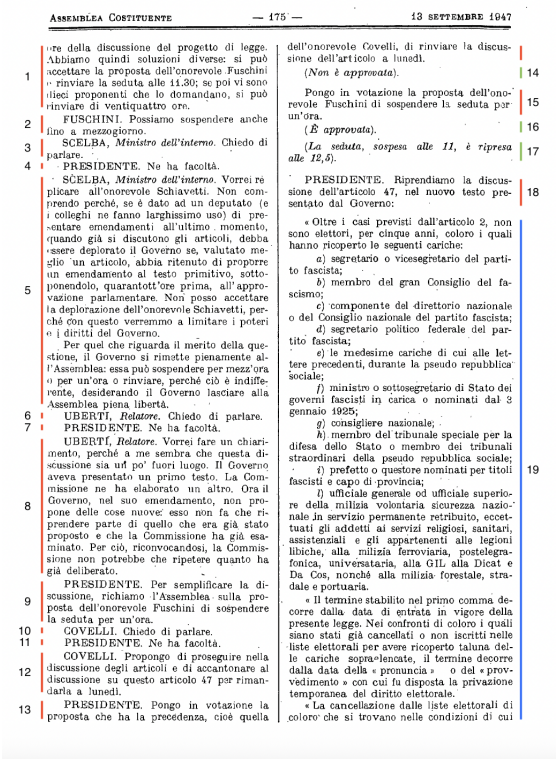
4. Annotated Dataset

Given the complexity of the annotated documents, which necessarily require detailed and articulated guidelines, an initial inter-annotator agreement phase was conducted on a randomly selected sample of 20 pages prior to proceeding with the annotation of the gold-standard data. Inter-annotator agreement (Table 2) was measured using Krippendorff's α for nominal data. Speaker attribution yielded substantial agreement ($\alpha = 0.76$), whereas agreement on type labels was considerably lower ($\alpha = 0.44$).

The discrepancy between the two scores reflects the different nature of the tasks. Speaker attribution is largely referential, whereas type label assignment requires interpretative distinctions between structurally adjacent categories (e.g., speech vs. description). The analysis of disagreements led to a refinement of the annotation guidelines before proceeding to the full gold-standard annotation, thereby improving overall consistency.

Following the agreement phase and the subsequent reconciliation of annotation discrepancies, we carried out a manual annotation on a randomly selected subset of the corpus comprising 300 pages, which were independently annotated by three annotators. Agreement was not recomputed on the full dataset, as the annotation process adopted an adjudication-based workflow aimed at producing a consistent gold-standard resource. More specifically, all disagreements concerning either type labels or speaker attribution were systematically identified and resolved through a joint adjudication process. Annotators collectively reviewed conflicting cases with reference to the annotation guidelines, refining their interpretation when necessary. This iterative reconciliation ensured consistency across the dataset and resulted in a fully harmonized gold standard.

Table 3 presents an example of the manual annotation of page 274 from the annotated subset, while Figure 1 provides a visual representation of the same page with color-coded and numbered boxes corresponding to the annotated segments (red for speech, green for description, and blue for text). Each box identifies a distinct functional block and is associated with its instance identifier, shown alongside the segment.


ASSEMBLEA COSTITUENTE — 175 — 13 SETTEMBRE 1947

re della discussione del progetto di legge. Abbiamo quindi soluzioni diverse: si può accettare la proposta dell'onorevole Fuschini o rinviare la seduta alle 11.30; se poi vi sono oltri proponenti che lo domandano, si può rinviare di ventiquattro ore.

FUSCHINI. Possiamo sospendere anche fino a mezzogiorno.

SCELBA, *Ministro dell'interno*. Chiedo di parlare.

PRESIDENTE. Ne ha facoltà.

SCELBA, *Ministro dell'interno*. Vorrei replicare all'onorevole Schiavetti. Non comprendo perché, se è dato ad un deputato (e i colleghi ne fanno larghissimo uso) di presentare emendamenti all'ultimo momento, quando già si discutono gli articoli, debba essere deplorato il Governo se, valutato meglio un articolo, abbia ritenuto di proporre un emendamento al testo primitivo, sottoponendolo, quarantott'ore prima, all'approvazione parlamentare. Non posso accettare la deplorazione dell'onorevole Schiavetti, perché con questo verremmo a limitare i poteri e i diritti del Governo.

Per quel che riguarda il merito della questione, il Governo si rimette pienamente all'Assemblea: essa può sospendere per mezz'ora o per un'ora o rinviare, perché ciò è indifferente, desiderando il Governo lasciare alla Assemblea piena libertà.

UBERTI, *Relatore*. Chiedo di parlare.

PRESIDENTE. Ne ha facoltà.

UBERTI, *Relatore*. Vorrei fare un chiarimento, perché a me sembra che questa discussione sia un po' fuori luogo. Il Governo aveva presentato un primo testo. La Commissione ne ha elaborato un altro. Ora il Governo, nel suo emendamento, non propone delle cose nuove: esso non fa che riprendere parte di quello che era già stato proposto e che la Commissione ha già esaminato. Per ciò, riconvocandosi, la Commissione non potrebbe che ripetere quanto ha già deliberato.

PRESIDENTE. Per semplificare la discussione, richiamo l'Assemblea sulla proposta dell'onorevole Fuschini di sospendere la seduta per un'ora.

COVELLI. Chiedo di parlare.

PRESIDENTE. Ne ha facoltà.

COVELLI. Propongo di proseguire nella discussione degli articoli e di accantonare al discussione su questo articolo 47 per rimandarla a lunedì.

PRESIDENTE. Pongo in votazione la proposta che ha la precedenza, cioè quella dell'onorevole Covelli, di rinviare la discussione dell'articolo a lunedì.

(*Non è approvata*).

Pongo in votazione la proposta dell'onorevole Fuschini di sospendere la seduta per un'ora.

(*È approvata*).

(*La seduta, sospesa alle 11, è ripresa alle 12,5*).

PRESIDENTE. Riprendiamo la discussione dell'articolo 47, nel nuovo testo presentato dal Governo:

« Oltre i casi previsti dall'articolo 2, non sono elettori, per cinque anni, coloro i quali hanno ricoperto le seguenti cariche:

a) segretario o vicesegretario del partito fascista;

b) membro del gran Consiglio del fascismo;

c) componente del direttorio nazionale o del Consiglio nazionale del partito fascista;

d) segretario politico federale del partito fascista;

e) le medesime cariche di cui alle lettere precedenti, durante la pseudo repubblica sociale;

f) ministro o sottosegretario di Stato dei governi fascisti in carica o nominati dal 3 gennaio 1925;

g) consigliere nazionale;

h) membro del tribunale speciale per la difesa dello Stato o membro dei tribunali straordinari della pseudo repubblica sociale;

i) prefetto o questore nominati per titoli fascisti e capo di provincia;

l) ufficiale generale od ufficiale superiore della milizia volontaria sicurezza nazionale in servizio permanente retribuito, eccettuati gli addetti ai servizi religiosi, sanitari, assistenziali e gli appartenenti alle legioni libiche, alla milizia ferroviaria, postelegrafonica, universitaria, alla GIL alla Dicat e Da Cos, nonché alla milizia forestale, stradale e portuaria.

« Il termine stabilito nel primo comma decorre dalla data di entrata in vigore della presente legge. Nei confronti di coloro i quali siano stati già cancellati o non iscritti nelle liste elettorali per avere ricoperto taluna delle cariche sopraelencate, il termine decorre dalla data della « pronuncia » o del « provvedimento » con cui fu disposta la privazione temporanea del diritto elettorale.

« La cancellazione dalle liste elettorali di coloro che si trovano nelle condizioni di cui

Figure 1: Visual representation of the annotation for page 274. Color-coded and numbered boxes indicate individual functional blocks, each labelled with its instance identifier, illustrating the segmentation process and its alignment with the document layout.

As can also be observed from the annotation example reported in 3, the annotation scheme is characterized by multiple layers of variation and complexity.

For instance, the same page includes different structural units, such as speech, description, and text, reflecting the heterogeneous nature of parliamentary proceedings. The majority of segments are labeled as speech, each associated with an explicit speaker attribution (e.g., Presidente, Scelba, Ministro dell'interno, Uberti, relatore), while other segments correspond to procedural descriptions (e.g., voting outcomes or session interruptions), which do not involve an identifiable speaker. The example also illustrates additional annotation decisions, such as the use of special markers for partially unavailable speaker information (e.g., [hidden]) or inferred attributions ([inferred] Presidente). This layered structure requires distinguishing between structural segmentation (ID-level units), discourse function (type labels), and speaker meta-

Label	Description
`speech`	Oral discourse delivered by MPs or institutional figures, typically introduced by a name. This includes interruptions such as *"Voci"* or *"Molte voci"*.
`speechnotext`	Instances where a speech act is referenced (e.g., a secretary reading a formal record) but the verbatim content is not provided in the transcript.
`article`	Legislative articles, usually introduced by the abbreviation "Art." followed by a numerical identifier.
`text`	Verbatim quoted written texts, official statements, or legal provisions read during a session. These are distinguished from oral speech via formatting such as indentation or angle quotes («. . . »).
`description`	Procedural notes regarding session management (e.g., opening/closing times) or standalone bracketed comments.
`presidencydeclaration`	Explicit indicators of the session's presidency, typically appearing in small caps below the title.
`title`	Structural headings for sessions, topics, or legislative proposals.
`other`	Non-discursive elements including signatures, tables, attachments, indexes, and content-related footnotes.

Table 1: Taxonomy of annotation labels for block types (label type)

Annotation task	Krippendorff's α
Speaker attribution	0.76
Type label classification	0.44

Table 2: Inter-annotator agreement measured using Krippendorff's α for nominal data on the initial 20-page sample.

Page	ID	Type	Speaker
274	1	speech	[hidden]
274	2	speech	Fuschini
274	3	speech	Scelba, Ministro dell'interno
274	4	speech	Presidente
274	5	speech	Scelba, Ministro dell'interno
274	6	speech	Uberti, relatore
274	7	speech	Presidente
274	8	speech	Uberti, relatore
274	9	speech	Presidente
274	10	speech	Covelli
274	11	speech	Presidente
274	12	speech	Covelli
274	13	speech	Presidente
274	14	description	
274	15	speech	[inferred] Presidente
274	16	description	
274	17	description	
274	18	speech	Presidente
274	19	text	

Table 3: Structured representation of page 274, including Page ID (Page), Instance ID (ID), type label (type), and speaker attribution (Speaker).

data, thereby increasing the overall annotation complexity.

Out of the 300 annotated pages, the complete set of type labels identified in the dataset is reported in Table 4.

Type label	Frequency	Percentage (%)
article	127	5.7
description	213	9.6
other	60	2.7
pres. declaration	11	0.5
speech	1585	71.2
speechnotext	22	1.0
text	89	4.0
title	121	5.4
Total	**2228**	**100.0**

Table 4: Distribution of type labels across the 300 annotated pages, with absolute and percentage frequencies.

Table 4 shows that the distribution of type labels is heavily skewed towards the speech category, which accounts for 71.2% of all annotated segments. This is expected given the dialogic and interactional nature of parliamentary proceedings, where the primary structural unit is the individual speech turn. At the same time, the presence of several additional categories, such as description (9.6%), article (5.7%), title (5.4%), and smaller classes like text, speechnotext, and presidency-declaration, highlights the structural heterogeneity of the documents. These categories capture procedural notes, editorial insertions, structural markers,

and non-speech material, which coexist with spoken interventions and may interrupt or frame them. Therefore, although speech segments clearly dominate the corpus quantitatively, the interaction and alternation between speech and non-speech units contribute significantly to the overall structural complexity of the annotation.

5. Experiments

The task of identifying and segmenting functional blocks in historical parliamentary debates is highly challenging. Unlike plain OCR transcription, this problem requires the reconstruction of the logical and communicative structure of the document.

Errors in reading order, missing typographical cues, or imperfect speaker markers can easily propagate into incorrect block segmentation and speaker attribution. Moreover, communicative boundaries are often signaled by subtle graphical features (e.g., indentation, capitalization, spacing) that may be partially lost during digitisation. For these reasons, block identification cannot be reduced to a simple text classification problem, but rather requires joint reasoning over layout, discourse function, and institutional conventions.

To evaluate the feasibility of automating this task, we experimented with two different approaches:

- Single-step direct block identification via LLMs
- Two-step pipeline: OCR followed by LLM-based block extraction

The experiments were conducted on the manually annotated corpus of 300 pages described in Section 4.

The prompt used for the LLM queries can be found in the Appendix.

5.1. Direct Block Identification

As a first baseline, we tested whether modern multimodal Large Language Models could directly perform block segmentation and speaker attribution in a single step.

In this configuration, the model was provided with:

- the annotation guidelines (in PDF format);
- the PDF page to be annotated;
- explicit instructions to segment the page into blocks following the guidelines.

This approach represents an upper-bound scenario in which the model is asked to jointly solve OCR, layout interpretation, functional classification, and speaker attribution within a single reasoning process.

However, preliminary observations show that this task is particularly demanding: performance degrades in pages containing dense legislative articles, embedded attachments, or unclear typographical separation between speech turns and procedural notes.

5.2. OCR + LLM-Based Block Extraction

Given the complexity of the single-step configuration, we designed a modular two-step pipeline that separates transcription from structural reconstruction.

5.2.1. Step 1: OCR

The first step consists of extracting raw textual content from scanned pages using different OCR systems. We evaluated the following engines: Mistral (multimodal OCR capabilities), Azure Document Intelligence, AWS Textract, Google Vision (Visual API), Tesseract.

The output of each OCR system consists of plain text (or structured text when available), which is then passed to the second stage. While these systems differ substrantially in terms of layout sensitivity and robustness to degraded scans and historical typography, Tesseract is the only one that is open source and available for free.

5.2.2. Step 2: LLM-Based Block Extraction

In the second step, the OCR output is provided to a Large Language Model tasked with:

- segmenting the text into functional blocks;
- assigning a type label from the predefined taxonomy;
- performing speaker attribution according to the annotation guidelines.

We evaluated the output of all the OCR systems with the following LLMs: Mistral Large, Gemini 3, GPT mini 5, Llama 4 Scout 17B 16e.

This modular configuration allows us to isolate the impact of transcription errors on structural reconstruction and the reasoning capabilities of different LLMs when operating on noisy OCR text.

5.3. Zero-Shot vs Few-Shot Configurations

The second approach (OCR + LLM extraction) was evaluated under two prompting strategies:

5.3.1. Zero-Shot Setting

In the zero-shot configuration, the model receives:

- the annotation guidelines;
- the OCR-extracted page text;
- instructions to produce the structured annotation.

No annotated examples are provided. This setting measures the model's ability to generalize directly from the schema description.

5.3.2. Few-Shot Setting

In the few-shot configuration, we provide the model with manually annotated examples before presenting the target page.

To ensure methodological rigor, the dataset of 300 annotated pages was split as follows:

- Development set (100 pages)
- Test set (200 pages)

Few-shot examples were sampled exclusively from the development set, while evaluation was conducted strictly on the held-out test set. No page in the test set was used as demonstration material.

The few-shot examples were selected to maximize structural diversity (e.g., pages dominated by speeches, pages with multiple legislative articles, presence of attachments, complex procedural notes).

5.4. Evaluation

The evaluation was designed to assess system performance along two complementary dimensions: (i) the correct identification and ordering of functional blocks, and (ii) the accurate transcription and attribution of speakers within speech segments.

The evaluation is performed on the 200 pages included in the test set (see Section 5.3.2).

5.4.1. Block Identification

For the first dimension, we evaluated the system's ability to reconstruct the correct sequence of annotated blocks on each page. For every test page, we generated two ordered sequences of labels:

- a gold sequence derived from the manually annotated corpus;
- a predicted sequence produced by the system.

Each element in the sequence corresponds to the type label assigned to a block (e.g., speech, description, article, text, etc.). Evaluation consists of measuring the distance between the gold and predicted sequences, taking into account insertions, deletions, substitutions, and ordering errors. This formulation allows us to capture both segmentation mistakes (e.g., merged or split blocks) and misclassification errors.

ChatGPT	Blocks	1338
	Speakers	639
Gemini	Blocks	715
	Speakers	373
Mistral	Blocks	1952
	Speakers	504

Table 5: Results of the single-step baseline (measured using Levenshtein distance).

5.4.2. Speaker Transcription and Attribution

The second dimension focuses specifically on speech segments. In this case, we constructed sequences composed only of blocks labeled as "speech" (see Section 3.1). For each speech block, the label corresponds to the extracted speaker string. Gold and predicted sequences were then compared after filtering out all non-speech elements. This setup evaluates both the correctness of speaker recognition and the ability to maintain the proper discourse order.

In both evaluation dimensions, correctness is measured using an edit-distance–based similarity metric grounded in Levenshtein distance. The metric is conceptually analogous to Word Error Rate (WER), which is commonly used to compare token sequences in speech recognition, and to its generalizations to higher-level semantic units, such as Slot Error Rate (SER). By operating on structured label sequences rather than individual tokens, the metric captures structural reconstruction quality rather than raw textual overlap (Ákos Tündik et al., 2020).

5.5. Results

The experimental results highlight clear performance differences across system configurations. Table 6 shows the edit-distance–based similarity metric grounded in Levenshtein distance. Table 5 shows the results for single-step approach (see next Section).

5.5.1. Baseline (Single-Step LLM)

The single-step direct block identification approach, in which the model jointly performs OCR, layout interpretation, classification, and speaker attribution, consistently yields the lowest performance across both evaluation dimensions. This confirms

			Mistral	Google	Tesseract	AWS	Azure
Llama 4	Zero-shot	Blocks	1418	1616	1566	1592	1368
		Speakers	468	666	696	752	604
	Few-shots	Blocks	1156	1060	636	970	652
		Speakers	464	636	570	630	418
GPT 5 mini	Zero-shot	Blocks	476	710	634	672	420
		Speakers	110	236	292	244	116
	Few-shots	Blocks	436	692	538	658	414
		Speakers	90	224	278	230	94
Gemini 3	Zero-shot	Blocks	546	756	562	546	498
		Speakers	96	202	76	98	56
	Few-shots	Blocks	580	560	388	400	338
		Speakers	88	210	92	66	56
Mistral large	Zero-shot	Blocks	1394	2410	1534	2038	2136
		Speakers	288	418	472	492	422
	Few-shots	Blocks	460	848	712	700	630
		Speakers	180	414	308	400	190

Table 6: Results of the accuracy (measured using Levenshtein distance) of the OCR + LLM approach (the lower, the better).

that solving OCR and structural reasoning simultaneously represents a particularly demanding task, especially in the presence of complex layouts and degraded historical scans.

5.5.2. Zero-Shot OCR + LLM

The modular two-step configuration (OCR followed by LLM-based block extraction) significantly improves results. In the zero-shot setting, performance varies depending on the specific combination of OCR engine and language model. The separation of transcription from structural reasoning allows LLMs to operate on textual input, reducing multimodal complexity. However, transcription noise from OCR systems still affects downstream segmentation and speaker attribution.

5.5.3. Few-Shot Configuration

The few-shot configuration achieves the best results for the block identification task, while achieving comparable results for speaker transcription. Providing annotated examples from the development set substantially improves structural alignment accuracy.

Interestingly, using OCR and LLM tools from the same provider (e.g., Mistral OCR + Mistral LLM, or Google Vision + Gemini) does not lead to systematic improvements. This suggests that performance is not determined by ecosystem coherence, but rather by the intrinsic quality of the transcription and the reasoning capabilities of the language model independently.

5.5.4. Language Model Comparison

Among the evaluated LLMs, Gemini consistently achieves the strongest performance across configurations and OCR inputs. It demonstrates robust handling of noisy OCR output and greater stability in maintaining correct block order and speaker attribution. Other models show higher variance depending on input quality and page complexity.

5.5.5. OCR System Comparison

Regarding transcription quality, Azure Document Intelligence emerges as the most reliable OCR system overall. Its outputs yield the best downstream structural reconstruction results, indicating superior robustness to historical typography and layout variability.

Notably, Tesseract, despite being open source and freely available, performs remarkably well. While it does not consistently surpass commercial systems, its results remain competitive, especially considering cost-effectiveness and reproducibility constraints.

6. Release

To foster transparency and reproducibility, we plan to publicly release both the annotated dataset and the annotation guidelines upon completion of the anonymized review process. An anonymized Google Drive folder has been shared with the re-

viewers.[6]

The release will include:

- the manually annotated corpus of 300 randomly sampled pages in structured format;
- the full annotation guidelines used during the project;
- the evaluation scripts required to reproduce the experimental results reported in this paper.

The dataset will be distributed in a machine-readable format designed to facilitate reuse. All resources will be made available under an open license compatible with research and academic use. By releasing both data and guidelines, we aim to provide a replicable benchmark for future work on historical parliamentary digitisation and to support the development of robust OCR and structural segmentation pipelines for long-span political corpora.

7. Conclusion and Future Work

This paper addressed the digitisation of historical Italian parliamentary proceedings (1848–1996), moving beyond plain OCR to tackle functional block segmentation and speaker attribution. We introduced tailored annotation guidelines and a manually annotated dataset of 300 randomly sampled pages.

We compared a single-step multimodal LLM approach with a modular OCR+LLM pipeline under zero-shot and few-shot settings. Results show that separating transcription from structural reasoning yields more reliable outputs, and that few-shot prompting significantly improves performance. Overall results are strongly influenced by OCR quality: Azure Document Intelligence provided the most robust inputs, Tesseract proved competitive as an open-source alternative, and Gemini achieved the best balance between textual accuracy and structural reconstruction across configurations.

Future work will focus on conducting additional experiments to identify the most robust and scalable configuration before applying the pipeline to the tens of thousands of pages of historical Italian parliamentary debates still awaiting structured digitisation. We also plan to evaluate the portability of both the annotation guidelines and the proposed pipeline to other languages and parliamentary traditions, assessing the degree of adaptation required across different institutional and layout conventions.

[6] https://drive.google.com/drive/folders/191ixt3e31EeTpa-ct36nzmiDxRmmY6wl

8. Limitations

The modular pipeline clearly demonstrates that structural reconstruction performance is strongly conditioned by transcription quality. OCR errors affecting capitalization, punctuation, indentation cues, or speaker markers propagate to downstream segmentation and attribution tasks.

Although Azure Document Intelligence performed best overall, commercial OCR systems introduce reproducibility and cost constraints. Conversely, while Tesseract remains competitive and reproducible, its sensitivity to degraded typography may limit scalability to lower-quality scans. Therefore, the overall framework remains critically dependent on the availability of high-quality OCR systems capable of handling historical Italian typography.

The annotation guidelines were designed specifically for the Italian parliamentary tradition and its layout conventions. Although the underlying principles of functional segmentation and speaker attribution are broadly applicable, direct transfer to other parliamentary corpora (e.g., Hansard, GERPARCOR, ParlaMint extensions) would likely require schema adaptation.

While the few-shot OCR+LLM pipeline achieves promising results, large-scale application to the entire parliamentary archive would entail:

- substantial computational costs for LLM inference;
- potential latency issues;
- and careful monitoring of error propagation across millions of pages.

Further optimization, model distillation, or hybrid rule-based post-processing may be required to ensure sustainable large-scale deployment.

9. Bibliographical References

Thomas M. Breuel. 2008. The OCRopus open source OCR system. In *Document Recognition and Retrieval XV*, volume 6815, page 68150F. International Society for Optics and Photonics, SPIE.

Gavin Greif, Niclas Griesshaber, and Robin Greif. 2025. Multimodal llms for ocr, ocr post-correction, and named entity recognition in historical documents.

Nuno Guimarães, Purificação Silvano, Ricardo Campos, Alipio Jorge, Ana Filipa Pacheco, Dimitar Iliyanov Dimitrov, Nikolaos Nikolaidis, Roman Yangarber, Elisa Sartori, Nicolas Stefanovitch,

Preslav Nakov, Jakub Piskorski, and Giovanni Da San Martino. 2025. NarratEX dataset: Explaining the dominant narratives in news texts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20408–20434, Suzhou, China. Association for Computational Linguistics.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. Ocr4all—an open-source tool providing a (semi-)automatic ocr workflow for historical printings. *Applied Sciences*, 9(22):4853.

Christoph Wick, Christian Reul, and Frank Puppe. 2018. Calamari - a high-performance tensorflow-based deep learning package for optical character recognition.

Máté Ákos Tündik, Balázs Tarján, and György Szaszák. 2020. A low latency sequential model and its user-focused evaluation for automatic punctuation of asr closed captions. *Computer Speech & Language*, 63:101076.

10. Language Resource References

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German parliamentary corpus (gerparcor). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1900–1906.

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German parliamentary corpus (gerparcor) reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7707–7716.

Francesca Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, and Alessandro Panunzi. 2022. Impaqts: un corpus di discorsi politici italiani annotato per gli impliciti linguistici. In *Corpora e studi linguistici = Corpora and linguistic studies: Atti del 54. Congresso internazionale di studi della Società di linguistica italiana (online, 8–10 settembre 2021)*, pages 151–164. Officinaventuno.

Matthew Coole, Paul Rayson, and John Mariani. 2020. Unfinished business: Construction and maintenance of a semantically tagged historical parliamentary corpus, UK Hansard from 1803 to the present day. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 23–27, Marseille, France. European Language Resources Association.

Jacopo Cova. 2025. A new database for italian parliamentary speeches: Introducing the itaparl-corpus dataset. *Italian Political Science Review / Rivista Italiana di Scienza Politica*, 55(1):77–86.

Luigi Curini, Silvia Decadri, Alfio Ferrara, Stefano Montanelli, Fedra Negri, and Francesco Periti. 2024. The gender gap in issue attention and language use within a legislative setting: An application to the italian parliament (1948–2020). *Politics & Gender*, 20(1):182–211.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, Sascha Diwersy, Maria Gavriilidou, 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 Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2024. Parlamint ii: Advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448.

Valentino Frasnelli and Alessio Palmero Aprosio. 2024. There's something new about the Italian parliament: The IPSA corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources*

and Evaluation (LREC-COLING 2024), pages 16037–16046, Torino, Italia. ELRA and ICCL.

Alenka Kavčič, Martin Stojanoski, and Matija Marolt. 2024. Historical parliamentary corpora viewer. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 127–132, Torino, Italia. ELRA and ICCL.

Guy Mor-Lan, Effi Levi, Tamir Sheafer, and Shaul R. Shenhav. 2024. IsraParlTweet: The israeli parliamentary and Twitter resource. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9372–9381, Torino, Italia. ELRA and ICCL.

Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. 2019. Semantifying the uk hansard (1918-2018). In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 412–413.

Ines Rehbein, Josef Ruppenhofer, Annelen Brunner, and Simone Paolo Ponzetto. 2024. Out of the mouths of MPs: Speaker attribution in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12553–12563, Torino, Italia. ELRA and ICCL.

Stephen Wattam, Paul Rayson, Marc Alexander, and Jean Anderson. 2014. Experiences with parallelisation of an existing NLP pipeline: Tagging Hansard. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4093–4096, Reykjavik, Iceland. European Language Resources Association (ELRA).

A. Prompts

We utilized differentiated prompting strategies for text extraction across the two experimental setups: a direct LLM-based extraction and a combined OCR-LLM pipeline. This ensured that the prompts were adapted to the unique input requirements of each method. The placeholders *{GUIDELINES}*, *{TEXT}*, *{FILE}* represent the task-specific instructions, the target input string and the source file, respectively.

A.1. LLM ONLY

Role: You are an expert archival researcher specializing in the digitization and annotation of Italian Parliamentary Debates. Your task is to segment a provided page of text into coherent "Textual Blocks" and label them according to strict linguistic and functional guidelines.

Core Task:

Analyze the provided page. Identify every distinct textual unit (Instance). For each unit, extract the required metadata into a CSV format.

Annotation Fields (CSV Columns):

1. **Page ID**: The integer page number.
2. **Instance/Round ID**: Progressive integer (1, 2, 3...) restarting at 1 for every new page.
3. **Type Label**: Must be exactly one of: `speech`, `speechnotext`, `article`, `text`, `description`, `presidencydeclaration`, `title`, `other`.
4. **Speaker**: The name/role of the speaker. Use `[unknown] Name` for generic voices, `[hidden]` if not printed, or `[inferred] Name` if clearly understood from context. Leave blank for `description` or `title`.
5. **Notes**: Brief description if Type is `other` (e.g., "signature", "table").

Classification Rules:

{GUIDELINES}

Constraint:

Return **ONLY** the CSV data. Do not provide introductory text or conversational fillers. Use a comma as a delimiter. Wrap text fields in double quotes if they contain commas.

Example Format:

```
Page ID,Instance/Round ID,Type label,
Speaker,Notes
12,1,presidencydeclaration,CASATI,
12,2,title,,Verificazione di poteri
12,3,speech,PRESIDENTE,
12,4,description,,(Approvato)
12,5,other,,signature
```

More instructions:

- Do not hallucinate extra rows. Only record distinct, visible structural blocks.
- A single 'speech' block usually contains multiple paragraphs. Do NOT create a new row for every paragraph or line; group them by the speaker.
- For this type of document, there are typically between 5 and 20 blocks per page. If you are exceeding 30 blocks, you are likely being too granular.
- If the speaker is unknown, do not guess; use '[unknown]'. However, do not use this to create repetitive filler rows.

{FILE}

A.2. OCR + LLM

Role: You are an expert archival researcher specializing in the digitization and annotation of Italian Parliamentary Debates. Your task is to segment a provided page of text into coherent "Textual Blocks" and label them according to strict linguistic and functional guidelines. The page is provided as an OCR transcription, therefore it can contain headers or footer that should be ignored.

Core Task:

Analyze the provided page. Identify every distinct textual unit (Instance). For each unit, extract the required metadata into a CSV format.

Annotation Fields (CSV Columns):

1. **Page ID**: The integer page number (it can be "X" if the number is not found/identified).
2. **Instance/Round ID**: Progressive integer (1, 2, 3...) restarting at 1 for every new page.
3. **Type Label**: Must be exactly one of: ‘speech‘, ‘speechnotext‘, ‘article‘, ‘text‘, ‘description‘, ‘presidencydeclaration‘, ‘title‘, ‘other‘.
4. **Speaker**: The name/role of the speaker. Use ‘[unknown] Name‘ for generic voices, ‘[hidden]‘ if not printed, or ‘[inferred] Name‘ if clearly understood from context. Leave blank for ‘description‘ or ‘title‘.
5. **Notes**: Brief description, not empty only if Type Label is ‘other‘ (e.g., "signature", "table").

Constraint:

Return **ONLY** the CSV data. Do not provide introductory text or conversational fillers. Use a comma as a delimiter. Wrap text fields in double quotes if they contain commas.

Example Format:

```
Page ID,Instance/Round ID,Type label,
Speaker,Notes
12,1,presidencydeclaration,CASATI,
12,2,title,,
12,3,speech,PRESIDENTE,
12,4,description,,
12,5,other,,signature
```

More instructions

- Do not hallucinate extra rows. Only record distinct, visible structural blocks.
- A single ’speech’ block usually contains multiple paragraphs. Do NOT create a new row for every paragraph or line; group them by the speaker.
- For this type of document, there are typically between 5 and 20 blocks per page. If you are exceeding 30 blocks, you are likely being too granular.
- If the speaker is unknown, do not guess; use ’[unknown]’. However, do not use this to create repetitive filler rows.

{GUIDELINES}

Could you extract the textual units from the text below, obtained through OCR?

{TEXT}