LocalGovPL: A Corpus of Speaker-Attributed Polish Local Government Transcripts

Dariusz Czerski, Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences
dariusz.czerski@ipipan.waw.pl, maciej.ogrodniczuk@ipipan.waw.pl

Abstract

We present LocalGovPL, a large-scale, speaker-annotated corpus of Polish local government meeting transcripts processed using an automatic two-stage LLM pipeline. The corpus consists of 31,900 sessions from 749 councils recorded between 2018–2025 (approximately 391M words). It is released in TEI P5 format with explicit links between utterances and registered participants. We collect transcripts from official local government portals using a dedicated crawler, normalize the text, and apply: (1) LLM-assisted extraction of person names and administrative roles; and (2) attribution of utterances to identified speakers using discourse cues. To evaluate attribution quality, we manually annotate 30 sessions and evaluate five LLM configurations using three evaluation protocols with speaker-aware word error rate (sWER). The strongest system, GEMINI-2.5-PRO, achieves 3.9% sWER for abstract speaker identification, 4.6% for known participants, and 5.9% for end-to-end processing with relaxed name matching. LocalGovPL enables large-scale analysis of local deliberative discourse and supports research on dialogue modeling, summarization, and political text analysis.

Keywords: local government transcripts, speaker attribution, Large Language Models

1. Introduction

In recent years, the increasing accessibility of public-sector transcripts has opened new possibilities for research on institutional communication, deliberative democracy, and political discourse. In particular, local government meetings, such as those of municipalities, counties, cities, and regional assemblies, constitute an important yet underexplored source of naturally occurring, domain-specific spoken interaction. However, despite their availability, these transcripts often remain unstructured and inconsistently formatted, which limits their usability for computational analysis and linguistic research. Initiatives such as the Council Data Project (CDP)[1] are designed to manually and semi-automatically collect and curate municipal governance data (Maxfield Brown and Weber, 2022).

A key challenge in transforming this data into a reusable linguistic resource is speaker attribution: identifying the speakers in a transcript and assigning utterances to them. This step is essential for downstream analyses, including dialogue modeling, stance detection, and discourse structure extraction. However, manual annotation of speaker information is extremely time-consuming and impractical given the scale and variability of available transcripts. Automated solutions are therefore crucial to enable large-scale processing of this type of data.

Recent advances in large language models (LLMs) offer new opportunities for addressing this challenge. LLMs are capable of interpreting complex textual cues and handling unstandardized inputs, making them suitable for tasks involving implicit structure recovery, such as identifying speakers and marking utterance boundaries in raw transcripts.

We present LocalGovPL (Ogrodniczuk and Czerski, 2025), a new speaker-attributed corpus of local government meeting transcripts retrieved from their respective official portals and curated using an automatic LLM-based processing pipeline. The pipeline consists of two stages: (1) extraction of potential speaker names, and (2) attribution of utterance segments to the identified speakers. To collect source transcripts at scale, we implemented a specialized web crawler that systematically retrieves and normalizes records from official government portals. We evaluated five LLM configurations across both subtasks, comparing their performance across several dimensions, including accuracy, consistency, and robustness to noise. The resulting corpus provides a structured representation of local-level deliberative discourse and constitutes a valuable resource for research in computational sociolinguistics, political text analysis, and dialogue processing.

The paper is structured as follows. Section 2 reviews related work on meeting transcript processing and speaker attribution. Section 3 describes data sources, processing pipeline, the corpus and

[1] https://councildataproject.org/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7883–7893
11–16 May 2026.

its statistics. Section 4 presents the LLM-based pipeline and its components. Section 5 details the evaluation methodology and experimental results. Section 6 discusses the implications and potential uses of the corpus, and Section 7 concludes with directions for future work and resource availability.

2. Related Work

Foundational meeting corpora, including AMI[2] and ICSI[3], underpin much work on transcription, diarization, and summarization (Janin et al., 2003; Carletta et al., 2006). Recent work has intensified the curation of municipal governance data, with the Council Data Project demonstrating an open infrastructure for assembling comparable local-government corpora (Maxfield Brown and Weber, 2022). Beyond data collection, specialized systems target key meeting understanding subtasks. For speaker identification from audio, Speakerbox[4] enables few-shot fine-tuning of transformer-based models to recognize previously seen speakers (Brown et al., 2023). Complementarily, PublicSpeak[5] detects and categorizes remarks made by constituents during public meetings using a probabilistic framework (Xu et al., 2025). A large body of research addresses automatic meeting summarization, including integrated transcription/summarization systems and both extractive and abstractive methods (Song et al., 2021; Vadlamudi et al., 2022; Martin, 2023; Tilkar, 2025). Complementary surveys in the government domain emphasize named entity recognition as a foundation for entity-level structuring (e.g. person names and roles) in institutional transcripts (Ramdhani et al., 2024).

A separate line of work explores text-based speaker attribution. Studies show that attribution based on linguistic content can be robust even when transcripts are produced automatically by ASR, highlighting stylistic and lexical cues as informative for identifying speakers (Aggazzotti et al., 2025). Pretrained language models have been used to assign utterances to a predefined set of participants using contextual dialogue cues (Nguyen et al., 2024). LLM-assisted labeling has also been used to map utterances to named speakers (Gobara et al., 2025; Yin et al., 2025).

At broader legislative levels, comparable parliamentary corpora such as ParlaMint[6] provide speaker- and role-annotated proceedings across many countries, offering design guidance for governance-oriented resources (Erjavec et al., 2023, 2024). On the acoustic side, speaker diarization has advanced through Bayesian HMM clustering of x-vectors (VBx[7]), end-to-end neural diarization (EEND[8]), and target-speaker voice activity detection (TS-VAD[9]), alongside broadly adopted toolkits such as pyannote.audio[10] (Fujita et al., 2019; Medennikov et al., 2020; Bredin et al., 2020; Landini et al., 2022).

Large language models have increasingly been used to improve the speaker diarization process and to couple diarization with recognition. In multi-talker scenarios, models combining Whisper[11] and WavLM encoders[12] with a small LLM fine-tuned via LoRA detect speaker changes under versatile instructions, though they do not attribute speech to named participants (Meng et al., 2025). As a post-processing step, LLMs can refine diarized transcripts and substantially reduce word diarization error rate (Wang et al., 2024), and fine-tuned models can correct diarization assignments with ensembles that generalize across different ASR systems (Efstathiadis et al., 2025). Beyond post-processing, unified speech LLMs jointly perform diarization and ASR end-to-end, showing strong results in multilingual, multi-speaker settings (Saengthong et al., 2025).

However, to our knowledge, these LLM-based diarization and attribution approaches have not been applied to local government meetings, where transcripts are heterogeneous, role-rich, and often noisy. Our work addresses this gap by applying LLMs to extract speakers and attribute utterances to named officials in municipal and regional proceedings.

3. Resource Description

3.1. Overview

The presented resource is a speaker-attributed corpus of local government meeting transcripts created through an automatic large language model LLM-based structuring pipeline. It includes transcripts from various levels of local administration – municipalities, counties, cities, and regional assemblies – covering both plenary sessions and committee meetings. The corpus provides a structured textual

[2] https://groups.inf.ed.ac.uk/ami/corpus/
[3] https://groups.inf.ed.ac.uk/ami/icsi/
[4] https://github.com/CouncilDataProject/speakerbox
[5] https://github.com/politechlab/publicspeak
[6] https://www.clarin.eu/parlamint
[7] https://github.com/BUTSpeechFIT/VBx
[8] https://github.com/hitachi-speech/EEND
[9] https://github.com/dodohow1011/TS-VAD
[10] https://github.com/pyannote/pyannote-audio
[11] https://github.com/openai/whisper
[12] https://huggingface.co/models?other=wavlm

representation in which utterance segments are attributed to identified speakers, enabling linguistic and computational analyses of political communication and deliberative discourse at the local level.

The primary goal of the resource is to facilitate research on the language of local governance, including studies of argumentation, interactional patterns, policy framing, and social dynamics within institutional dialogue. Beyond linguistic research, the corpus supports applications in speech-to-text alignment, automatic summarization, speaker role identification, and computational social science.

3.2. Data Sources

The raw transcripts were collected from two main publicly available sources: websites maintained by local administrative bodies and the meeting streaming platform used by local governments[13]. In the first case, a set of specialized HTML extraction parsers was implemented. In the second case, transcription files in the WebVTT format were downloaded. The dataset covers meetings from 2018 to 2025 and includes several thousand hours of deliberation.

Due to the decentralized publication practices of local institutions, the source transcripts exhibit substantial variability in format, structure, and language conventions. The preprocessing stage therefore included normalization of document encoding, removal of metadata irrelevant to the spoken discourse (e.g., agenda headers or timestamps), and segmentation into individual utterance candidates.

3.3. Data Processing Pipeline

The automatic structuring pipeline consisted of two main stages: speaker extraction and utterance attribution.

3.3.1. Speaker Extraction

Potential speaker names were identified using a combination of rule-based name recognition and contextual inference performed by LLMs. The models were prompted to detect person names and administrative roles (e.g., Chairperson, Mayor, Council Member), ensuring both high recall and accurate disambiguation in cases of title repetition or partial name mentions.

3.3.2. Utterance Attribution

The LLMs were then used to assign each utterance segment to one of the previously extracted speakers. This stage required interpreting discourse cues such as addressing forms, transitions, and speaker introductions. The output of this stage was a fully structured transcript, in which each utterance is associated with a speaker identifier and metadata (speaker name, role, and meeting session).

[13] https://esesja.tv/

```
Analyze the following transcription and extract all unique speakers mentioned or implied in
the text.
Transcription: transcription
Instructions: 1. Identify all unique speakers in the transcription 2. Look for explicit speaker
mentions, titles, roles, or contextual clues 3. Return only the speaker names/identifiers,
one per line, templates like: prezydent <name> <surname> pani skarbnik <name>
<surname> radna <name> <surname> radni <name> <surname> <name> <surname>
<title>
4. Try to identify speakers by their name and/or surname, combine title/role with
name/surname 5. Do not include any explanations or additional text 6. If no speakers are
clearly identifiable, return an empty list
Speakers:
```

Figure 1: Speaker extraction prompt.

```
Task: Identify Speaker Changes in a Polish Transcription
Instructions:
Analyze the transcription and identify speaker changes. Output the row_id and speaker
name in the following format: <row_id>ẑspeaker name>
Use context, speaking patterns, and content to detect speaker transitions.
Do not include explanations or any additional text in your output.
Speaker Change Cues (Templates):
Use the following common patterns as indicators of a speaker change:
0. Start of Transcription – Unnamed Speaker If the transcription begins without a named
speaker, it is most likely the Moderator/Chairperson (Przewodniczący posiedzenia).
1. Direct Introduction by the Moderator/Chairperson The current speaker (usually the
chairperson) introduces the next speaker.
Examples:
Proszę / Poproszę / Zapraszam o zabranie głosu [Tytuł/Stanowisko] [Imię Nazwisko].
Oddaję głos [Tytuł/Stanowisko].
O (odczytanie / przedstawienie) [Nazwa Dokumentu] proszę [Tytuł/Stanowisko].
[Nazwa Dokumentu] przedstawi [Tytuł/Stanowisko] [Imię Nazwisko].
2. Self-Introduction or Formal Address by a New Speaker A new speaker begins with a
formal greeting or self-introduction, often marked by a dash ("-").
Examples:
- Panie/Pani Przewodniczący, Wysoka Rado, Szanowni Państwo.
- Dziękuję. / Dziękuję bardzo. (when it begins a new turn)
3. Moderator Transition / Conclusion The chairperson ends a speech and signals the
return of control or the next step.
Examples:
Dziękuję bardzo, [Tytuł/Stanowisko].
Dziękuję. Przechodzimy do (następnego punktu / głosowania).
Dziękuję. Czy są pytania do...?
5. If a line begins with a '-', it most likely indicates a speaker change.
Important:
Only output speaker information for lines where a speaker change occurs.
Do not tag every line.
Available speakers:
Transcription (format: <row_id>ẑtext>) (polish language): transcription
Continue until you reach the final line of the transcription. last_line
Speaker changes:
```

Figure 2: Utterance attribution prompt.

3.3.3. Processing Configuration Used for the Released corpus

For the public release of LocalGovPL, both stages were executed end-to-end with DEEPSEEK-CHAT-V3-0324. We used the concise prompts shown in Figures 3.3.3 and 3.3.3 and the change-only output format described in Section 4. Long transcripts were processed with the chunking strategy (threshold $> 1{,}500$ lines of transcription, approximately 60,000 characters) and merged by global line numbers as detailed below. This configuration was chosen to balance accuracy, throughput, and cost at corpus scale; as reported in Section 5, GEMINI-2.5-PRO attains lower sWER and can be used to reprocess subsets that demand near-gold attribution.

3.3.4. Throughput and Cost at Corpus Scale

Running the end-to-end pipeline over the full collection of 31,900 transcripts consumed a total of 1,100,000,000 input tokens and 55,000,000

output tokens. The end-to-end processing took 16.82 days. API charges amounted to $373.18 in total, which corresponds to $0.01078 per transcript on average. On average, each transcript required 34,038.3 input tokens and 1,742.3 output tokens, and the API reported an average generation time of 41.964 s per request. Table 1 summarizes these figures.

Table 1: Processing throughput and cost for producing the released corpus.

Metric	Value
Transcripts processed	31,900
Total input tokens	1,100,000,000
Total output tokens	55,000,000
Total processing time (days)	16.82
Total cost (USD)	373.18
Avg input tokens per transcript	34,038.3
Avg output tokens per transcript	1,742.3
Avg generation time (s)	41.964
Avg cost per transcript (USD)	0.01078

3.4. Corpus Format and Structure

The corpus is released in TEI P5 XML format and follows the same design choices as the Polish Parliamentary Corpus (PPC; Ogrodniczuk, 2018), ensuring interoperability with tools and queries described in (Ogrodniczuk and Nitoń, 2020). Each meeting transcription is represented by a pair of XML files: `header.xml` and `text_structure.xml`.

3.4.1. Session Header

The `header.xml` file contains the TEI header with document-level metadata and the participant registry. In detail, it includes meeting metadata:

- `title` – meeting title used as the document name (e.g., *Sesja Rady 30 stycznia 2019* (EN: *Council Session on January 30, 2019*))
- `publisher` – the organizing body responsible for the session (e.g., *Rada Miejska Nowego Miasta Lubawskiego* (EN: *Municipal Council of Nowe Miasto Lubawskie*))
- `system` – source system label for provenance tracking (e.g., *Sesja Rady Lokalnej* (EN: *Local Council Session*))
- `house` – assembly or chamber type (e.g., *Rada Powiatu* (EN: *County Council*))
- `sitting ID` – numeric identifier of the sitting
- `type` – content type of the source (e.g., *Transkrypcja sesji* (EN: *Session transcript*))
- `total rows` – number of input transcript rows prior to structuring
- `speaker count` – number of distinct speakers recognized in the session
- `date` – session date in ISO format (e.g., *2019-01-30*)

and participant metadata (each `person` is uniquely identified and carries a normalized name and role):

- `structure` – `person[@xml:id]` provides a stable identifier (e.g., `chairman_of_municipal_council`); `persName` holds the display name (e.g., *Przewodniczący Rady Miejskiej*; EN: *Chairman of the Municipal Council*); an implicit or explicit role is encoded via `@role` and/or role-bearing text within `persName` (e.g., *Burmistrz Gminy*; EN: *Mayor of the Municipality*)
- `uniqueness` – `xml:id` values are unique within the header and serve as targets for utterance references.

3.4.2. Utterance Structure

The `text_structure.xml` file contains the speech content segmented into `<div>`isions and `<u>`tterances. Each utterance carries:

- `xml:id` – a unique utterance identifier (e.g. `u-1.1`)
- `who` – a pointer to the speaking participant using a TEI cross-reference to `header.xml`
- `start`/`end` – timestamps delimiting the utterance span in the source recording.

Documents may be wrapped in a `<teiCorpus>` element that includes `header.xml` via XML Inclusions (`xi:include`); regardless of wrapping, the logical linkage between utterances (`<u>/@who`) and declared speakers (`<listPerson>/person[@xml:id]`) remains the same. This representation mirrors PPC conventions to enable reuse of existing tooling and facilitate cross-corpus comparisons.

3.5. Corpus Statistics

The LocalGovPL corpus represents a substantial collection of local government meeting transcripts, spanning over seven years of administrative proceedings.

Table 2 presents the detailed statistics of the corpus, highlighting its extensive scope and diversity.

The corpus consists of about 391 million words and 2.21 billion characters of transcribed content. Each session averages approximately 12,250

Table 2: Detailed statistics of the corpus

Category	Count	Average per Session
Basic Statistics		
Total transcripts	31,900	–
Date range	2018-11 to 2025-08	–
Number of councils	749	–
Transcripts per council	–	42.59
Duration Statistics		
Average session duration	–	2.23 hours
Content Statistics		
Total words	390,777,715	12,250
Total characters	2,206,514,586	69,170
Speaker Statistics		
Average speakers per session	–	12.77
Average utterances per session	–	62

words and 69,170 characters. The speaker participation patterns reveal active multi-participant discussions, with an average of around 12.8 speakers per session contributing approximately 62 utterances each.

The temporal coverage spans from November 2018 to August 2025. The geographic distribution across 749 councils ensures representative coverage of diverse administrative contexts and regional variations in governance practices.

4. LLM-based Pipeline

The pipeline has two stages: speaker extraction and utterance attribution. We use short, instruction-style prompts shown in Figures 3.3.3 and 3.3.3. We kept the language of the prompts simple and task-focused to improve robustness across heterogeneous transcripts.

4.1. Stage 1: Speaker Extraction

In the first stage, the model scans the transcription to collect potential speakers. It returns names together with roles or titles when available (for example, *przewodniczący* (EN: *chairman*), *burmistrz* (EN: *mayor*), *skarbnik* (EN: *treasurer*), *radny* (EN: *councilor*)). We deduplicate and normalize the list. Producing this list in advance has two benefits: it reduces the search space for the second stage and helps the model use consistent speaker names later on, even when introductions are implicit or abbreviated.

4.2. Stage 2: Utterance Attribution

In the second stage, the model receives (a) the list of available speakers from Stage 1 and (b) a compact set of speaker-change cues (Figure 3.3.3). The transcription is provided as numbered lines. The model is asked to output only those lines where the speaker changes, in the format `<row_id>\t<speaker_name>`. This reduces the required output length and helps preserve the original transcription.

We observed that attribution is harder than name collection because it requires reading local context and discourse markers (e.g., formal greetings, explicit handovers, and moderator transitions). Giving the model both the candidate speaker list and clear cues improves accuracy and consistency.

To avoid overgeneration, the utterance-attribution prompt ends with an explicit end-of-transcript instruction and a terminal sentinel (rendered as `{last_line}`, Figure 3.3.3). We include this instruction because we observe that LLMs may continue producing outputs even after the provided transcription ends; the instruction and sentinel explicitly bind generation to the final line of the input.

4.3. Input and Output Format

The model reads the transcription as simple tab-separated lines (see Figure 3). The system only outputs speaker attribution lines when the speaker changes, not for every line. This compact format allows the system to reconstruct continuous speaker segments by assigning each line to the most recent

Input (format: `<row_id>\t<text>`):

```
1  Dzień dobry
2  Nazywam się Jan Nowak i będę
   przewodniczył dzisiejszej sesji.
3  Chciałbym oddać głos
   Panu Janowi Kowalskiemu.
4  Dziękuję, Panie Przewodniczący.
5  Przedstawię Państwu sprawozdanie
   z poprzedniej sesji.
```

Output (only on speaker changes; format: `<row_id>\t<speaker name>`):

```
1 Jan Nowak
4 Jan Kowalski
```

Figure 3: Input and output format for the utterance attribution stage. (Polish example with English translation: "Good morning" / "My name is Jan Nowak and I will chair today's session" / "I would like to give the floor to Mr. Jan Kowalski" / "Thank you, Mr. Chairman" / "I will present to you a report from the previous session")

speaker until the next change line. This approach avoids having the model rewrite the full transcript, reducing the risk of omissions or unintended edits.

4.4. Handling Long Transcripts

Local government meetings often exceed two hours, and WebVTT transcripts frequently exceed 2,000 lines. Full prompts can easily approach or exceed large context windows, which slows inference and may degrade quality. For very long inputs (more than 1,500 lines), we process the transcription in chunks. Each chunk is formatted identically and the model again outputs only speaker-change lines. Because we number the full transcription up front, we can safely merge chunk-level outputs by line number to recover a single, consistent list of changes for the whole session.

4.5. Design Choices and Rationale

We originally asked the model to produce a full JSON mapping from lines to speakers. This approach proved inefficient and sometimes led to changed or skipped content. Switching to change-only outputs made the task lighter and more reliable. Numbering lines lets us rely on stable indices rather than fragile text matching when building the final, structured transcript. Finally, separating extraction (Stage 1) from attribution (Stage 2) keeps prompts short and lets the model focus on a narrow decision at each step.

5. Evaluation

5.1. Test Dataset

A subset of 30 transcripts was manually annotated to create a reference benchmark for evaluating both speaker identification and attribution. Human annotators verified the correctness of speaker name extraction and the alignment between utterances and speakers.

Each session lasts approximately 2.4 hours and contains nearly 15,000 words, with an average of 17 speakers contributing about 94 utterances per session.

5.2. Speaker Identification (Stage 1)

We first assess Stage 1, which identifies the set of unique speakers present in a session and establishes the participant inventory used downstream. To compare predicted and reference speaker sets, we apply the relaxed identity equivalence defined in Section 5.3.4. We report macro-averaged precision, recall, and F1 over our 30-session benchmark (Section 5.1). Table 3 summarizes the results across model configurations.

Overall, larger-capacity models achieve the strongest macro scores (GEMINI-2.5 ≈ 0.88 macro F1), while smaller/open models lag. These outcomes indicate that Stage 1 is strong enough to supply reliable candidate participant lists, enabling a focused evaluation of utterance attribution in Stage 2.

5.3. Speaker Attribution (Stage 2)

5.3.1. Evaluation Setup and Metrics

We evaluated five LLM configurations across three complementary evaluation protocols and report speaker-aware word error rate (sWER; lower is better). More precisely, sWER extends standard word error rate by requiring that a hypothesized token is counted as correct only if both its lexical form and its speaker label match the reference. Let the reference be a sequence of labeled tokens (w_i, s_i) and the system output (w'_j, s'_j). We compute the minimum-edit alignment over these word-speaker pairs and report $\mathrm{sWER} = (S + D + I)/N$, where N is the number of reference words; S counts substitutions whenever $w_i \neq w'_j$ or $s_i \neq s'_j$, D are deletions, and I are insertions under this alignment. Thus, words attributed to the wrong speaker are penalized as substitutions even when the lexical content is correct. Below we describe the considered protocols.

5.3.2. Abstract Speaker Attribution

This protocol evaluates the system's ability to distinguish between different speakers without requiring

Table 3: Speaker identification (Stage 1) macro metrics (averaged over 30 sessions; relaxed identity equivalence per Section 5.3.4).

Model configuration	Macro P	Macro R	Macro F1
gemini-2.5-pro	0.9058	0.8814	0.8786
gemini-2.5-flash	0.9071	0.8800	0.8783
deepseek-chat-v3-0324	0.8287	0.8375	0.8169
deepseek-r1-0528	0.6281	0.5887	0.5904
llama-3.3-70b-instruct	0.3537	0.3673	0.3491

exact name matching. Speakers are treated as abstract entities (e.g., `speaker-1`, `speaker-2`) rather than specific individuals. This approach focuses purely on the core challenge of determining *when* the speaker changes in a conversation, regardless of whether the system correctly identifies *who* is speaking.

For example, if a transcript contains three speakers (*Burmistrz Kowalski* (EN: *Mayor Kowalski*), *Radny Nowak* (EN: *Councilor Nowak*), and *Skarbnik Wiśniewski* (EN: *Treasurer Wiśniewski*)), the system might label them as `speaker-A`, `speaker-B`, and `speaker-C`. We then use the Hungarian algorithm to find the optimal one-to-one mapping between the system's abstract speakers and the reference speakers, ensuring that each system speaker is matched to exactly one reference speaker. This protocol isolates the utterance attribution task from name recognition challenges, providing a baseline for how well the system can detect speaker changes in the conversation flow.

5.3.3. Ground-truth Participants

This protocol tests the system's utterance attribution performance when given perfect information about who the speakers are. We bypass Stage 1 (speaker extraction) entirely and provide the system with the gold-standard list of actual participants from the meeting. The system then only needs to determine which of these known speakers is talking at each point in the transcript.

This approach isolates the utterance attribution task from speaker identification errors. For instance, if we know the meeting participants are "*Burmistrz Anna Kowalska*" (EN: "Mayor Anna Kowalska"), "*Radny Jan Nowak*" (EN: "Councilor Jan Nowak"), and "*Skarbnik Maria Wiśniewska*" (EN: "Treasurer Maria Wiśniewska"), the system receives this exact list and must only decide which of these three people is speaking at each moment. This protocol answers the question: "Given perfect knowledge of who is present, how accurately can the system determine who is speaking when?"

5.3.4. End-to-end with Relaxed Name Matching

This protocol evaluates the complete pipeline (both speaker extraction and utterance attribution) under realistic conditions where the system must handle the full complexity of the task. Both stages run end-to-end without any external assistance, simulating real-world deployment where the system must independently identify speakers and attribute utterances.

To account for the natural variation in how names appear in transcripts, we use relaxed matching criteria. A predicted speaker is considered a match to the reference if: (a) surnames match (e.g., "Kowalski" matches "Kowalski"); (b) titles/roles match (e.g., "*Burmistrz*" (EN: "Mayor") matches "*Burmistrz*" (EN: "Mayor")); or (c) the Levenshtein similarity between names is at least 0.8 (allowing for minor spelling variations or abbreviations). This approach reflects real-world scenarios where names might appear in different forms (e.g., "Jan Kowalski" vs. "J. Kowalski" vs. "*Burmistrz Kowalski*" (EN: "Mayor Kowalski")) but refer to the same person.

This protocol provides the most realistic assessment of system performance, as it evaluates the complete pipeline under conditions that mirror actual deployment, where the system must handle both the complexity of identifying speakers from context and the challenge of matching names across different formats and variations.

5.3.5. Results

We report averages across 30 manually annotated sessions for each protocol. Across the three protocols, GEMINI-2.5-PRO (Comanici et al., 2025) is the strongest model; the abstract-speaker protocol yields the lowest error rates overall, with the ground-truth-participants protocol close behind; the end-to-end relaxed setting is slightly more challenging. Table 4 reports average sWER values.

Overall, these results indicate that higher-capacity proprietary models achieve the strongest attribution accuracy across all protocols, while some open models struggle with strict output formatting and long-context discourse cues.

Table 4: Speaker attribution sWER (average across 30 sessions; lower is better) under three evaluation protocols.

Model configuration	Abstract	GT participants	Relaxed names
gemini-2.5-pro	0.0393	0.0460	0.0592
gemini-2.5-flash	0.0907	0.1287	0.1257
deepseek-chat-v3-0324	0.2061	0.2094	0.2381
deepseek-r1-0528	0.4582	0.2498	0.4684
llama-3.3-70b-instruct	0.6969	0.7945	0.7378

6. Discussion

The released LocalGovPL corpus was produced using the DEEPSEEK-CHAT-V3-0324 configuration across both stages of the pipeline (speaker extraction and utterance attribution). Among the three evaluation protocols (abstract speakers, ground-truth participants, and end-to-end with relaxed names), abstract attribution attains the lowest sWER, with the ground-truth-participants evaluation method close behind and the relaxed end-to-end protocol showing higher error rates (Table 4). This indicates that current LLMs can reliably detect speaker changes and maintain consistent speaker identities, while precise name extraction is the primary source of remaining errors.

High-capacity models perform best in all protocols: GEMINI-2.5-PRO is strongest (0.0393 abstract; 0.0460 ground-truth participants; 0.0592 relaxed), whereas some open models are less robust, often due to strict formatting and long-context constraints. Despite these differences, the present corpus quality is sufficient for many linguistic and political-science analyses that emphasize aggregate patterns rather than per-utterance perfection.

We observe several recurring challenges typical for local-government transcripts and LLM-based structuring:

- Name variation and role confusion: morphological variation (e.g., case inflection in Polish) and repeated titles can cause occasional confusion between participants with similar names or roles.
- Boundary ambiguity: short responses or formulaic politeness markers at turn edges can shift boundaries by one line, especially around "Dziękuję"/"Proszę" exchanges.
- Noisy or templated source lines: remaining agenda items, captions, or system-inserted boilerplate can be spuriously attributed if not fully filtered during preprocessing.

These issues are infrequent at the corpus scale and therefore have minimal practical consequences.

7. Conclusion and Future Work

This paper presented LocalGovPL, a large-scale, speaker-attributed corpus of local government meeting transcripts structured with a two-stage LLM pipeline. The resource spans 2018–2025 across 749 councils and 31,900 sessions, and is released in interoperable TEI P5 XML format with explicit links between utterances and registered participants. Owing to its size and structure, the corpus supports a broad range of research tasks, including computational sociolinguistics, discourse and interaction analysis, dialogue modeling, summarization, speaker role identification, and political text analysis. For Stage 1 (speaker identification), large models achieved the best macro scores (GEMINI-2.5 ≈ 0.88 macro F1), providing reliable candidate lists for Stage 2.

In future work, we will integrate automatic speaker diarization over the source audio and use the resulting time-aligned speaker turns to constrain and correct text-based attribution. We also plan to exploit on-screen captions and nameplates that appear in some video recordings: OCR of these overlays can provide weak but valuable supervision for resolving speaker identities and roles, especially when introductions are implicit or abbreviated.

Beyond these directions, we plan to train and evaluate a compact model specialized for this task via supervised fine-tuning and distillation from higher-capacity models. In particular, we plan to use the change-only labels and participant registries produced by GEMINI-2.5-PRO (or the current best model) as pseudo-gold to fine-tune a smaller instruction model, aiming to approach top-tier attribution accuracy while reducing cost and latency for routine updates. We also plan a full-corpus refresh with GEMINI-2.5-PRO to improve overall quality. Based on current batch pricing, we estimate an end-to-end cost roughly $3\times$ that of DEEPSEEK-CHAT-V3-0324; given the observed sWER gains, we consider this suitable for periodic releases.

The LocalGovPL corpus is available for research and educational use from the project website[14]. Our crawler updates the corpus regularly as new

[14] https://zil.ipipan.waw.pl/LocalGovPL

meetings are published, and refreshed releases will incorporate these additions. We welcome feedback and community collaboration on extensions and downstream benchmarks.

8. Ethical Considerations

All data used in this study originate from official public records published by governmental institutions. The collection and redistribution of these materials is conducted in compliance with the Polish Act of 11 August 2021 on Open Data and the Re-use of Public Sector Information[15], which mandates the openness of public sector information for reuse. The corpus does not include any personal data beyond names of public officials acting in their professional capacity. No manual modifications were made to the linguistic content of the transcripts.

Risk of Misattribution. As an automatically processed resource, the corpus may contain attribution errors (sWER $\approx$ 4–6%). While high-capacity models demonstrate strong performance, misattributions can occur, particularly in cases of rapid speaker turns or implicit introductions. Users should exercise caution when attributing specific controversial or sensitive statements to individual public officials based solely on this automated dataset.

The resource is intended solely for research and educational purposes, and all derivative uses must comply with applicable open-data regulations.

9. Acknowledgments

This work was financed as part of the investment: CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01 and by CLARIN-PL, the European Regional Development Fund, FENG programme, agreement number FENG.02.04-IP.040004/24.

10. Bibliographical References

Cristina Aggazzotti, Matthew Wiesner, Elizabeth Allyn Smith, and Nicholas Andrews. 2025. The Impact of Automatic Speech Transcription on Speaker Attribution. *Transactions of the Association for Computational Linguistics*, 13:1578–1596.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pages 7124–7128. IEEE.

Eva Brown, To Huynh, and Nicholas Weber. 2023. Speakerbox: Few-Shot Learning for Speaker Identification with Transformers. *Journal of Open Source Software*, 8:5132.

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, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In Steve Renals and Samy Bengio, editors, *Proceedings of International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005), Lecture Notes in Computer Science, vol 3869*, pages 28–39. Springer, Berlin, Heidelberg.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Georgios Efstathiadis, Vijay Yadav, and Anzar Abbas. 2025. LLM-based speaker diarization correction: A generalizable approach. *Speech Communication*, 170:103224.

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*, 59:2071–2102.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin,

[15] Dz.U. 2021 poz. 1641

Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, de Macedo Luciana D., Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darǵis, Ruben Ring, Orsolya van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019. End-to-end Neural Speaker Diarization with Self-Attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303. IEEE.

Seiji Gobara, Hidetaka Kamigaito, and Taro Watanabe. 2025. Speaker Identification and Dataset Construction Using LLMs: A Case Study on Japanese Narratives. In *Proceedings of The 7th Workshop on Narrative Understanding*, pages 97–119.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI Meeting Corpus. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, volume 1. IEEE.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.

Paul Martin. 2023. Meeting Summarizer using Natural Language Processing. *International Journal for Research in Applied Science and Engineering Technology*, 11:188–195.

Eva Maxfield Brown and Nicholas Weber. 2022. Councils in Action: Automating the Curation of Municipal Governance Data for Research. *Proceedings of the Association for Information Science and Technology*, 59(1):23–31.

Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko. 2020. Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario. In *Proceedings of Interspeech 2020*, pages 274–278.

Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng. 2025. Large Language language Model Can Transcribe Speech in Multi-Talker Scenarios with Versatile Instructions. In *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*, pages 1–5. IEEE.

Minh Nguyen, Franck Dernoncourt, Seunghyun Yoon, Hanieh Deilamsalehy, Hao Tan, Ryan Rossi, Quan Hung Tran, Trung Bui, and Thien Huu Nguyen. 2024. Identifying Speakers in Dialogue Transcripts: A Text-based Approach Using Pretrained Language Models. In *Proceedings of Interspeech 2024*, pages 3799–3803.

Maciej Ogrodniczuk. 2018. Polish Parliamentary Corpus. In *Proceedings of the LREC 2018 Workshop* ParlaCLARIN: Creating and Using Parliamentary Corpora, pages 15–19, Paris, France. European Language Resources Association (ELRA).

Maciej Ogrodniczuk and Bartłomiej Nitoń. 2020. New Developments in the Polish Parliamentary Corpus. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 1–4, Marseille, France. European Language Resources Association (ELRA).

Tosan Wiar Ramdhani, Indra Budi, and Betty Purwandari. 2024. Named entity recognition in government domain: A systematic literature review. *Journal of Infrastructure, Policy and Development*, 8:9789.

Phurich Saengthong, Boonnithi Jiaramaneepinit, Sheng Li, Manabu Okumura, and Takahiro Shinozaki. 2025. A Unified Speech LLM for Diarization and Speech Recognition in Multilingual Conversations. In *Proceedings of the Workshop on Multilingual Conversational Speech Language Model (MLC-SLM)*, pages 56–60.

Yuanfeng Song, Di Jiang, Xuefang Zhao, Xiaoling Huang, Qian Xu, Raymond Chi-Wing Wong, and Qiang Yang. 2021. SmartMeeting: Automatic Meeting Transcription and Summarization for In-Person Conversations. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 2777–2779, New York, NY, USA. Association for Computing Machinery.

Swati Tilkar. 2025. A Review on Natural Language Processing (NLP) Models for Generating Meeting Transcription. *International Jornal of Scientific Research Engineering and Management*, 9:1–9.

Geethika Vadlamudi, Naveena Vemuru, Surendhra Vangapalli, Ravi Kishan Surapaneni, and Sailaja

Nimmagadda. 2022. Meeting Summarizer using Natural Language Processing. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1610–1614.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. 2024. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models. In *Proceedings of Interspeech 2024*, pages 3754–3758.

Tianliang Xu, Eva Maxfield Brown, Dustin Dwyer, and Sabina Tomkins. 2025. PublicSpeak: hearing the public with a probabilistic framework. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. 2025. SpeakerLM: End-to-End Versatile Speaker Diarization and Recognition with Multimodal Large Language Models. arXiv:2508.06372.

11. Language Resource References

Ogrodniczuk, Maciej and Czerski, Dariusz. 2025. *LocalGovPL corpus*. Institute of Compute Science, Polish Academy of Sciences, 1.0. PID https://zil.ipipan.waw.pl/LocalGovPL.