ParlaMint-TZ: ParlaMint-ready Corpus of Tanzanian Hansards

Beata Wójtowicz
University of Warsaw
b.wojtowicz@uw.edu.pl

Maciej Ogrodniczuk
Michał Rudolf
Dariusz Czerski
Institute of Computer Science
Polish Academy of Sciences
first_name.surname@ipipan.waw.pl

Abstract

This paper presents ParlaMint-TZ, a standardized parliamentary corpus constructed from Tanzanian Hansards spanning 2003 to 2025, alongside select archival records from 1975. Comprising 1,642 unstructured PDF documents (nearly 300,000 pages) and approximately 90 million words, primarily in Swahili with occasional English, this resource enables systematic, large-scale research on parliamentary discourse, political communication and multilingual practices. To overcome structural heterogeneity, inconsistent speaker naming and non-speech annotations, an open-source, deterministic rule-based processing pipeline was developed using ParlaMint-PL architecture to extract metadata and structure proceedings into ParlaMint-compliant TEI XML format. Finally, the corpus is integrated into an online platform featuring full-text search capabilities, term highlighting and structured transcript rendering, bringing African parliamentary data into a globally interoperable research ecosystem.

1 Introduction

The ParlaMint corpus (Erjavec et al., 2023, 2025) provides multilingual, comparable parliamentary data spanning 29 European countries and autonomous regions, encompassing over one billion words. This format relies on a consistent and strictly defined TEI XML schema enforcing uniform encoding for both content and rich metadata concerning speakers (including party affiliation, roles and functions), as well as the structural organization of parliamentary sessions (such as headings, transcriber annotations and non-verbal incidents).

Although one existing study evaluated the feasibility of encoding non-European (South African) parliamentary proceedings into the standardized ParlaMint TEI schema (Ogrodniczuk, 2024), it used only a single, randomly selected English-language session transcript, serving as a small-scale pilot experiment. Applying the ParlaMint schema to Tanzanian data is intended to be the first experiment on a much larger scale.

The primary contribution of this paper is a documented, rule-based workflow for transforming Tanzanian parliamentary Hansards (*Bunge*) from heterogeneous PDF documents into a ParlaMint-compatible TEI XML corpus. The resulting corpus constitutes the main data resource, while the web platform provides a means of accessing and searching the processed material. Beyond presenting the resource itself, the paper identifies domain-specific challenges in speaker identification, structural segmentation and the interpretation of parenthetical material, offering practical insights for extending ParlaMint to parliamentary records from other African contexts.

2 Source Data

2.1 Dataset Description and Related Work

The project aims to create a consistent and unified text corpus of Tanzanian Parliamentary Proceedings (Hansards), which are freely available online through the Parliament of Tanzania's website[1].

The current collection comprises 1,642 unique documents covering the period from 2003 to 2025. Together, these documents amount to nearly 300,000 A4 pages in PDF format and approximately 90 million words, making the collection a substantial resource for the study of parliamentary discourse, language use, political communication and socio-political change in Tanzania. In addition to the contemporary material, the collection includes two archival files documenting parliamentary proceedings from 6–8 November 1975. Although these archival materials are limited in scope, they provide an important historical reference point and may contribute to extending the temporal coverage of the corpus in the future.

[1] https://www.parliament.go.tz/hansard

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

The Tanzanian Parliament currently meets four times a year, with each parliamentary session comprising multiple sittings – each one noted in separate documents. The resulting Hansards document debates, speeches, interventions and other forms of parliamentary discourse, offering a detailed record of political and institutional communication over time. Swahili is the primary language of parliamentary debate and the official working language of the Parliament, while English is also occasionally used, particularly in individual statements or specific communicative contexts. The bilingual nature of the material makes the corpus particularly relevant for research on multilingual parliamentary discourse and language practices in Tanzania.

The project will therefore provide a structured and standardized corpus that brings together parliamentary materials from different periods into a single research-oriented resource. Aligning the collection with the ParlaMint schema will facilitate interoperability with other parliamentary corpora and enable comparative research across countries, languages and political systems.

To date, Tanzanian parliamentary Hansards have been used in a number of linguistic and discourse-oriented studies, including research on language choice and code-switching (Dzahene-Quarshie, 2011), as well as parliamentary argumentation and discourse by (Msagalla and Visser, 2021) and (Msagalla, 2020). However, previous studies have generally relied on selected transcripts, individual parliamentary debates, or limited periods rather than on a comprehensive, consistently structured corpus covering the parliamentary record over an extended period. A limited selection of Tanzanian Hansards from 2004–2006 is also included in the Helsinki Corpus of Swahili 2.0. The present project therefore seeks to bring together a substantially larger body of material and make it available in a standardized, interoperable format suitable for systematic corpus-based research.

2.2 Document Layout and Pre-processing

The data preparation process involved the automated extraction of both textual content and metadata, including the date of the session, sitting number and starting and ending times. Although the documents share a broadly recognizable structure, variations in metadata, speaker labels and annotations required additional processing during corpus preparation.

The files have a clear and well-organized structure. The main challenges arose in the identification and normalization of speakers and in the interpretation of various types of non-speech and parenthetical information embedded in the proceedings.

For the purposes of the corpus, texts from the relevant part of the proceedings were included. These texts generally appear under headings such as *Maswali na Majibu* (‘Questions and Answers’), *Taarifa ya Spika* (‘Speaker’s Statement’), *Maswali kwa Waziri Mkuu* (‘Questions to the Prime Minister’), *Mwongozo wa Spika* (‘Speaker’s Guidance’), etc., which had to be recognised. The preamble usually contains information about the prayer and a list of documents laid before Parliament, i.e. *Hati zilizowasilishwa mezani* (’Documents Submitted/ Laid on the Table’). These documents appear after the question-and-answer section and also take the form of discussions. The main body of the proceedings is divided into numbered questions and answers, each provided with a title, like for example *Na. 53 Ujenzi Kwenye Maeneo ya Wazi Mji wa Himo* (’No. 53: Construction on Open Land in the Town of Himo’).

2.3 Speaker Normalisation and Structural Heterogeneity

Speakers are identified in a highly variable manner. In some cases, they are represented by an honorific title, such as *Mhe.* (from *Mheshimiwa*, an honorific title roughly equivalent to ‘Honourable’), followed by their full name, as in *MHE. AUGUSTINO M. MASELE*. In other cases, the name is preceded by an additional title or professional designation, such as *DKT.* (*Daktari*, ‘Doctor’) or *KAPT.* (*Kapteni*, ‘Captain’). The speaker’s official position or institutional role may then be provided after the name. However, speakers are not always identified by their personal names. In some cases, only their official position is given, for example *WAZIRI WA NCHI, OFISI YA RAIS (UTUMISHI)*, while in other instances the speaker is identified only generically as *MBUNGE FULANI* (‘a certain Member of Parliament’).

An additional complication arises when a person speaks on behalf of another individual or institution and is identified accordingly. For example, a speaker may be represented as *WAZIRI WA NCHI, OFISI YA RAIS (UTUMISHI) (K.n.y. OFISI YA WAZIRI MKUU)*, indicating that the statement is made on behalf of another office or institution. The relationship between the speaker’s name and their

official position is also expressed in different ways. The name may precede the position, follow it after a dash, or appear in parentheses. Examples include *NAIBU WAZIRI WA FEDHA – MHE. ADAM K. MALIMA* and *NAIBU WAZIRI WA NISHATI NA MADINI (MHE. CHARLES M. KITWANGA)*. Consequently, speaker identification requires more than simple recognition of personal names: it involves distinguishing between names, honorifics and professional titles, parliamentary or governmental positions and information indicating that a person is speaking on behalf of another individual or institution.

A further difficulty is the inconsistent spelling of personal names. The same individual may appear in the Hansards with variations in the spelling of their first name, surname, or both. Such inconsistencies may involve differences in capitalization, orthography, or the representation of particular names and they can make it difficult to determine whether differently written forms refer to the same person. This is particularly problematic when attempting to establish a consistent speaker inventory and to link all interventions by a given individual across different sittings and sessions. Consequently, speaker identification may require not only the recognition of titles and institutional roles, but also the comparison and normalization of variant name forms. The normalization procedure should therefore distinguish between the extraction of the original speaker label, the identification of its constituent elements (e.g. honorifics, personal names and institutional positions), and the assignment of a normalized speaker identifier where the available evidence supports such a decision. Ambiguous cases should be retained for manual review or explicitly marked as unresolved rather than automatically merged.

Another challenge involved the identification of comments and other types of non-speech information embedded in the parliamentary proceedings. These include indications of audience reactions, such as *makofi* ('applause') and *kicheko* ('laughter'), which typically appear in parentheses within the text. Such elements need to be distinguished from the actual spoken content and appropriately represented in the corpus.

The use of parentheses, however, is not restricted to comments or audience reactions. Parenthetical material is also used for a variety of other purposes, including abbreviations and their expansions, alternative forms of names, institutional or organizational names and additional explanatory or clarifying information. As a result, the identification of parenthetical material cannot be based solely on its formatting. Instead, its function has to be determined from the context in order to distinguish genuine comments and audience reactions from editorial or explanatory information and material that forms part of the speaker's utterance.

3 Data Processing Pipeline

The processing pipeline developed for transforming the transcripts of the Parliament of Tanzania into the ParlaMint format employs a traditional, rule-based approach. The conversion framework leverages existing code architecture from the ParlaMint-PL pipeline combined with tailored pattern recognition heuristics which deterministically identify, segment and label standard elements within the source transcript text, including speaker attributions, procedural events, metadata notes and spoken utterance contents.

While Large Language Models (LLMs) have demonstrated strong capabilities in general text transformation tasks, their application to exact, schema-compliant TEI/XML conversion for low-resource parliamentary corpora presents distinct structural limitations such as lack of precise schema and task-specific knowledge, absence of domain-specific training data or inability to infer unseen structural conventions. Three key examples from the Tanzanian parliamentary proceedings illustrate these domain-specific requirements:

1. In the source transcripts, parliamentary questions begin with a numbered heading followed by the question body. During conversion, these elements must be consolidated into a single paragraph and transformed into a formal `<note>` element representing the question metadata.

2. Spoken utterances frequently contain explicit delegation indicators, such as: *MHE. OLIVER D. SEMUGURUKA K.n.y. MHE. ERIC J. SHIGONGO aliuliza*. The rule-based conversion must explicitly disentangle the proxy statement from the primary speaker attribution.

3. In addition to structural transformations, parliamentary processing requires robust handling of formatting variations and typographical errors in structural headers (e.g., session identifiers with misspellings).

While an LLM can easily generate regular expressions for canonical pattern matches, it fails to handle unpredicted typographical variations in an automated pipeline. Because LLM generation is isolated from execution, it cannot adapt dynamically during runtime processing. Addressing such issues with LLMs would require either preprocessing text normalization steps (where automated OCR/spelling correction may still introduce errors) or extensive manual corrections.

In contrast, our deterministic pipeline utilizes a controlled feedback loop: uncaptured or misaligned files are automatically flagged during batch conversion, allowing target heuristics to be iteratively updated or specific edge cases to be corrected manually. This ensures full compliance with the ParlaMint standard while maintaining greater control over data integrity.

4 Browsing and Searching the Corpus

The corpus has been made available via a lightweight web interface with full-text search capabilities and structured transcript rendering[2].

The system's user interface features a persistent top navigation header providing direct access to the main platform modules: the full-text search engine (`Search`) and the document archive (`Documents`).

4.1 Full-Text Search Engine

The search subsystem allows users to perform queries across the corpus. Figure 1 shows the search interface after executing a query for the Swahili word *alichagua* ("he/she chose").

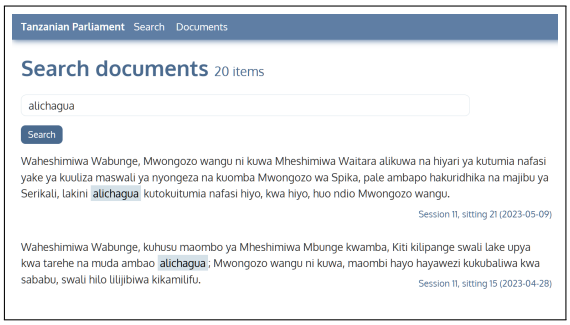

Figure 1: The full-text search interface displaying query results, hit counters, highlighted terms in context and metadata links to individual sittings.

The search view incorporates several key features to facilitate efficient information retrieval:

- **Hit counter** displays the total number of matching entries found in the database (e.g. *"20 items"*).
- **Key-Word-in-Context (KWIC)** view presents contextual text fragments around matched query terms, enabling users to evaluate relevance without opening full documents.
- **Visual term highlighting** of query terms matching the search string.
- **Session hyperlinking** for each result entry (to the corresponding position on the transcript page).

4.2 Structured Transcript View

The document archive shows an index of sittings and allows for rendering the transcript of an individual sitting. As illustrated in Figure 2, the interface delineates speakers, procedural headers and dialogue.



Figure 2: The structured transcript view displaying speaker turns, parliamentary questions and the official debate.

The transcript rendering module preserves essential structural elements of the parliamentary debate:

- **Session and sitting header** with macro-level metadata, including the session number, sitting number and date.
- **Roles and individual speaker names** are styled as section headers (in Fig. 2: *Mwenyekiti* for the Chairperson and *Mhe. Oran M. Njeza* for Members of Parliament, where *Mhe.* abbreviates *Mheshimiwa* / Honorable).
- **Content markup** with formal question numbers, constituency references and responses from government officials.

[2] See `http://parlamint-tz.ipipan.waw.pl/`

5 Conclusions and Further Steps

This project presents a large-scale workflow for transforming Tanzanian parliamentary Hansards from unstructured PDF documents into a standardized ParlaMint-compatible corpus, offering new opportunities for large-scale research on parliamentary discourse, language use, political communication and socio-political change in Tanzania.

A central contribution of the project is the development of an open-source processing pipeline that combines automated metadata extraction, formatting-aware PDF parsing, speaker and speech-turn identification and conversion to ParlaMint TEI XML. Particular attention was required for the normalization of speaker information and the interpretation of parenthetical material, which proved to be among the most heterogeneous aspects of the source documents. The resulting representation preserves distinctions between speakers, institutional roles, spoken content and procedural or non-verbal information while providing a consistent structure suitable for further computational processing.

The use of the ParlaMint schema also makes the Tanzanian material interoperable with other parliamentary corpora. This is particularly important for comparative research, as it enables Tanzanian parliamentary discourse to be studied alongside parliamentary data from other countries and linguistic contexts. At the same time, the corpus provides a valuable resource for research on Swahili, including studies of multilingual parliamentary discourse, code-switching, political vocabulary, argumentation and changes in language use over time.

Several aspects of the resource remain open for further development. In particular, more comprehensive speaker normalization and linking of variant speaker names could improve the consistency of the speaker inventory. Morphological annotation would also substantially increase the usefulness of the corpus for linguistic research. However, the lack of a suitable openly licensed morphological analyzer currently limits this possibility. In the meantime, regular-expression-based search and other corpus-query techniques can provide a practical solution for investigating selected morphological patterns in the highly agglutinative Swahili language.

Several technical limitations should be acknowledged. The conversion of parliamentary documents from PDF into structured TEI XML presents challenges related to document layout, speaker identification and the consistent representation of parliamentary proceedings. Some ambiguities in the source documents may therefore remain unresolved. These limitations should be taken into account when interpreting and reusing the corpus and highlight the need for further improvements to the conversion workflow and validation procedures.

The reuse and redistribution of the source documents and derived corpus data also depend on the applicable copyright and licensing conditions. The availability of documents on the Parliament's website does not necessarily imply permission to redistribute them. We are currently working to clarify the copyright and licensing conditions governing the use and redistribution of the source documents and the derived corpus data. The applicable conditions should be documented clearly once they have been established.

Future work will therefore focus on extending the linguistic annotation of the corpus, improving speaker normalization and developing more advanced search and analysis functionality. The availability of the processing pipeline and the standardized corpus representation should also facilitate the extension of the resource to additional parliamentary years and, potentially, to other African parliamentary corpora. In this respect, the project demonstrates that the ParlaMint framework can serve as a practical basis for bringing parliamentary data from underrepresented linguistic and geographical contexts into a broader ecosystem of interoperable parliamentary language resources.

Acknowledgments

The submission was supported by 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.

References

Josephine Dzahene-Quarshie. 2011. Language Policy, Language Choice and Language Use in the Tanzanian Parliament. *Legon Journal of the Humanities*, 22:27–69.

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, and 18 others. 2025. 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, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, and 9 others. 2023. The ParlaMint Corpora of Parliamentary Proceedings. *Language Resources and Evaluation*, 58:415–448.

Brighton Msagalla. 2020. *Argumentation involving account-giving and self-presentation in Tanzanian parliamentary debates: A pragma-dialectical perspective*. Ph.D. thesis. Stellenbosch University, South Africa.

Brighton Msagalla and Marianna Visser. 2021. Agenda-setting through Topic Shift in Tanzanian Parliamentary Debate: The Derailment of Strategic Manoeuvring. *Southern African Linguistics and Applied Language Studies*, 39(4):327–337.

Maciej Ogrodniczuk. 2024. Towards Including South African Hansard Papers in the ParlaMint Schema. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).