Towards Including South African Hansard Papers in the ParlaMint schema

Ogrodniczuk, Maciej
Institute of Computer Science
Polish Academy of Sciences
maciej.ogrodniczuk@gmail.com

Abstract

The ParlaMint project, a CLARIN flagship initiative, seeks to standardize the representation of parliamentary data across diverse languages and regions. Version 3.0 of ParlaMint encompasses corpora from 26 European countries and autonomous regions, available for download and search under the CC-BY license. These corpora adhere to a common XML encoding schema, ensuring interoperability. This study evaluates the feasibility of applying the ParlaMint schema to the proceedings of the Parliament of the Republic of South Africa. Through the conversion of a randomly selected parliamentary session, we scrutinize how various elements are modelled, delineating the steps required to initiate a comprehensive encoding endeavour.

The experiment starts with data retrieval by downloading Hansard records from the South African Parliament website. An English session was selected to streamline processing for non-South African researchers. The original format consisted of session headers, metadata, introductory messages, and debate records. Speeches were identified by uppercase headers and segmented into paragraphs.

Transcript conversion entailed extracting data from the PDF, eliminating technical elements, and ensuring continuity of utterances. Speaker names and functions were identified, and the session was transformed into ParlaMint-compliant TEI XML format. Meta-comments, including applause, laughter, and interruptions, were categorized based on typical phrases.

Quotations, marked with indentation in the original transcript, were manually encoded as TEI elements. Foreign-language fragments were treated as gaps, with English translations provided. Multiparagraph foreign utterances were encoded paragraph by paragraph.

Speaker metadata was stored in separate XML files, listing organizations and individual speakers. Speaker names and roles were converted into XML IDs, and web pages were linked for additional information. Speaker type was designated based on metadata, distinguishing between chairs, guests, and regular speakers.

The encoded session comprised 95 utterances, with varying distributions among speakers. The proposed conversion process serves as a starting point for the larger endeavour of encoding South African parliamentary data in the ParlaMint schema. While not exhaustive, this study lays the groundwork for expanding the ParlaMint dataset to include African parliamentary records.

Keywords: parliamentary debates, parliamentary records, parliamentary corpora, ParlaMint

1 Introduction

Parliamentary debates possess unique content, structure, and language, making them significant subjects of study in fields like political science, sociology, history, discourse analysis, sociolinguistics, and information technology. The CLARIN research infrastructure has played a longstanding role in organizing work around parliamentary data. The development of previous recommendations informed the development of ParlaMint (Erjavec et al. 2023) — a CLARIN flagship project aimed at harmonizing the representation of parliamentary data across languages. ParlaMint in its current version (Erjavec et al. 2023*a*,*b*, Kuzman et al. 2023) contains corpora for 26 European countries and autonomous regions, openly available under the CC-BY license for download and search.

The corpora are intended to be interoperable by encoding them according to a common ParlaMint

schema. The following paper intends to test the applicability of the ParlaMint schema to the proceedings of the Parliament of the Republic of South Africa. During the conversion of one randomly selected parliamentary session we analyse how various phenomena are modelled and propose the steps which need to be taken to start a full-scale encoding project.

2 Methodology

Based on previous work in ParlaMint, the detailed methodology of adding new corpora to the dataset was compiled. Most parts of the process are performed locally, i.e. by a project partner interested in adding their data to the ParlaMint infrastructure. The integration steps are carried out by the ParlaMint technical managers.

1. First, the parliamentary data and metadata are acquired by the interested party (most likely, a research organisation interested in creating the corpus). It may involve various methods, depending on the availability and format of the original data, e.g. scraping it from the parliamentary websites, obtaining it via parliamentary or third-party API or even retrieving it from an already maintained parliamentary corpus.
2. Then, the data is converted into the ParlaMint schema. It can be performed in many ways, also depending on the format, e.g. from HTML to basic TEI XML (TEI Consortium 2017) and then to the ParlaMint format, through XSLT stylesheets or by writing own scripts with heuristics for difficult parts.
3. Data validation (formal and qualitative) is performed on various levels. First, validation of corpus samples with ParlaMint scripts is performed locally. Secondly, samples are incorporated into the ParlaMint GitHub repository where external validation scripts are run. Only then the complete corpus can be processed and submitted. The validation procedure involves both basic TEI compatibility and detailed ParlaMint-dependent constraints applied using specialized RelaxNG schemata and XSLT scripts.
4. Linguistic processing is performed in-house (by each partner, most frequently using their own language-dependent tools) or by applying a Universal Dependencies-based (Marneffe et al. 2021) morphosyntactic annotation with named entities of several categories (person names, locations, organisations, other).
5. Machine translation of the corpus into English is performed externally, by the ParlaMint technical managers. It is intended to provide researchers with the opportunity to search across all available ParlaMint corpora, in all languages. For this purpose, the pre-trained OPUS-MT models (Tiedemann & Thottingal 2020) are used with additional proper name corrections.
6. The corpus is made available for download via the CLARIN.SI repository and for search through the noSketch Engine (Kilgarriff et al. 2014) and KonText (Machálek 2020) concordancers. Its translated version is similarly mounted on the concordancers as a parallel corpus.
7. The documentation of the corpus is created using the template common for all ParlaMint corpora, with sections describing general corpus information, information on data sources, the process of data acquisition and the method used for linguistic annotation.

3 Data retrieval

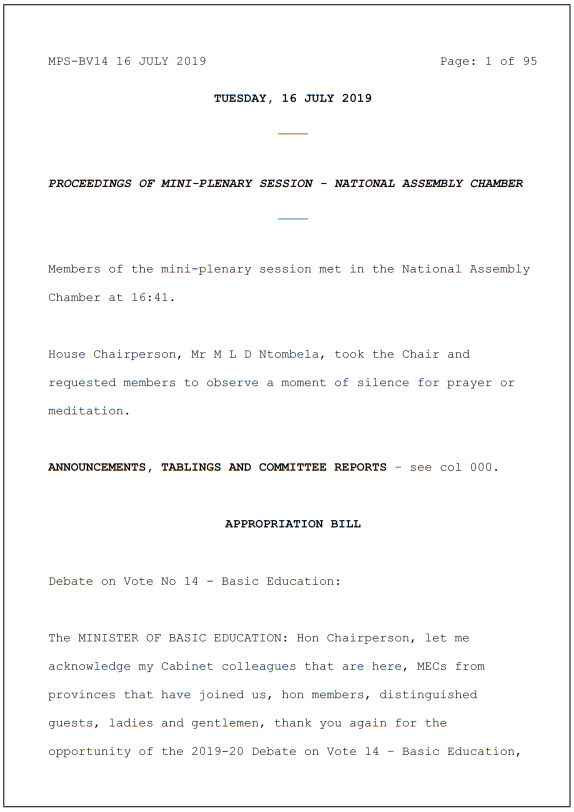
The Parliament of South Africa is bicameral and comprises a National Assembly (400 seats) and a National Council of Provinces (90 seats). The Hansard of the sessions of both houses ("a substantially verbatim report — with repetitions and redundancies omitted and obvious mistakes corrected — of parliamentary proceedings" (Parliament of The Republic of South Africa 2023*e*) are available online and can be used to "view, copy, download to a local drive, print and distribute the content of this

website, or any part thereof for informational or reference purposes only and for non-commercial purposes" (Parliament of The Republic of South Africa 2023*c*).

South African Hansard records can be downloaded from the *Hansard papers* section (Parliament of The Republic of South Africa 2023*d*) of the South African Parliament website. The sample session was selected by looking up the most recently produced one at the time of the experiment (August 8, 2023). Even though the proceedings are published in various official languages of South Africa: Ndebele, Pedi, Sotho, Swati, Tsonga, Tswana, Venda, Xhosa, Zulu, and English (most of the records and all after 2015), the selected session data was in English which facilitated further processing by a non-South African researcher. At the same time, a quick look at other language Hansard reports showed the same structure as English which makes the current experiment applicable to other session data.

The selected file (`HAN-MPS-BV14-2019-07-16.pdf`) represented a session from July 16, 2019, and was available in the PDF format of 95 pages.

Other sources of the same session were also consulted: one by the Parliamentary Monitoring Group (Parliamentary Monitoring Group 2023) and another one by the People's Assembly website (Parliament of The Republic of South Africa 2023*a*) but they occurred derivative.



MPS-BV14 16 JULY 2019 Page: 1 of 95

TUESDAY, 16 JULY 2019

PROCEEDINGS OF MINI-PLENARY SESSION - NATIONAL ASSEMBLY CHAMBER

Members of the mini-plenary session met in the National Assembly Chamber at 16:41.

House Chairperson, Mr M L D Ntombela, took the Chair and requested members to observe a moment of silence for prayer or meditation.

ANNOUNCEMENTS, TABLINGS AND COMMITTEE REPORTS – see col 000.

APPROPRIATION BILL

Debate on Vote No 14 - Basic Education:

The MINISTER OF BASIC EDUCATION: Hon Chairperson, let me acknowledge my Cabinet colleagues that are here, MECs from provinces that have joined us, hon members, distinguished guests, ladies and gentlemen, thank you again for the opportunity of the 2019-20 Debate on Vote 14 - Basic Education,

Figure 1: The first page of the retrieved transcript.

4 The original format

Each page of the file contained a header with the session identifier, date, page number and the total number of pages in the file: `MPS-BV14 16 JULY 2019 Page: 1 of 95`.

The file contained some metadata and introductory messages (see Fig. 1) which were followed by the record of the debate as a series of speeches. After the closing remarks about adjourning the session a link to "Announcements, Tablings and Committee Reports" for the day was included.

Each speech has the form of a header representing the speaker (name, function or both), output in uppercase, e.g. *The HOUSE CHAIRPERSON (Mr M L D Ntombela)* which is separated from the content with a colon. The speech text is split into paragraphs.

Meta-comments are given in square brackets. They can correspond to kinesic events such as `[Applause.]` but also regular comments such as `[Time expired.]`.

The text can also contain foreign-language fragments, uttered without notification, as part of the speech. When a longer fragment is spoken in a language different from the main language of the document, it is marked in parentheses, e.g. "Translation of isiZulu paragraph follows.)". The translation is given in square brackets.

Shorter fragments may not be marked at all, as in `Mamelani. [Listen.]`

5 Transcript conversion

The data were first extracted from the PDF file and saved in plain text format. Technical introductory notes and page headers were removed and since they could appear in the middle of an utterance, a simple heuristic was used: when the first utterance on a page was starting in lowercase, it was glued to the preceding utterance from the previous page, forming one continuous paragraph.

Since speaker names or functions are output in uppercase, the session was split into turns by looking at at least two uppercase fragments followed by a colon. Individual paragraphs forming one turn were extracted and output into target XML in the ParlaMint format — utterances modelled as <u> elements containing separate <seg>-ments with speakers identified in who attribute:

```
<u who="#The_MINISTER_OF_BASIC_
        EDUCATION"
  xml:id="ParlaMint-ZA_HAN-MPS-BV14-
        2019-07-16.u0"
  ana="#guest">
  <seg xml:id="seg1">Hon Chairperson,
    let me acknowledge my Cabinet
    colleagues that are here,
    ...
  </seg>
</u>
```

Meta-comments (together with some distinctive pieces of text, not marked separately as meta-comments) were converted into Parla-CLARIN elements according to their function, using simple heuristics, mostly based on typical phrases used in the text (see Table 2).

One issue to point out here is the distinction between notes about time when the meeting started and adjourned and the information about time expiration. Initially the latter had been modelled as time notes but eventually the decision was made to model them as vocal interruptions. It was motivated by the presence of other-than-English equivalents of the phrase "Time expired" which may suggest that it was uttered by the chair rather than added as a comment by the stenographer.

Apart from these removals, the text was not modified in any other way (including punctuation) which may help reproduce the original minutes from the ParlaMint-compliant XML.

6 Quotations

Quotations were marked in the original transcript with an indentation which was difficult to detect automatically. Since the ParlaMint schema does not allow the TEI <quote> element to appear in <seg>ments, the two quotations present in the session transcript were encoded manually as <note>s with a non-standard type quote:

```
<seg xml:id="seg237">
  I quote Phuti Seloba:
  <note type="quote">I can put my head
    on the block and say that all the
    outstanding toilets will be
    delivered before the end of 2014.
    We just need our people to be
    patient with us!</note>
```

7 Other language fragments

Fragments in other languages are marked as gaps, with the English translation available as the main content. This is in line with ParlaMint guidelines (Erjavec et al. 2023) which motivate this decision in the following way:

> Sometimes a passage of the transcription is in a foreign language, and, especially as the corpus is to be linguistically annotated, the passage is best left out of the transcription proper. This can be achieved by encoding it as a gap in the transcription with the reason foreign, while the <desc> should contain the omitted text.

This way the content:

> Sifisa ukuncoma nokuwubonga uMnyango ngalolu hlelo oluhle abalwenzela izwe lakithi eNingizimu Afrika.

Table 1: Heuristics used to convert meta-comments into Parla-CLARIN element types

Event	Type / Reason	Typical phrases
note	time	at HH:MM (events with time expression)
	debate	debate on..., took the chair
	quote	quotation marked with block indentation (see Section 6)
kinesic	applause	applause
	laughter	laughter
vocal	clarification	It is a point of debate.
	interruption	time expired
	exclamation	interjections
gap	inaudible	inaudible
	foreign	translation... follows

(Translation of isiZulu paragraph follows.)

[We wish to applaud and thank the department for this programme that they have put in place for our country, South Africa.]

becomes represented as:

```
<gap reason="foreign">
  <desc xml:lang="zu">Sifisa ukuncoma
  nokuwubonga uMnyango ngalolu hlelo
  oluhle abalwenzela izwe lakithi
  eNingizimu Afrika.</desc>
</gap>
<seg xml:id="seg224">We wish to
  applaud and thank the department
  for this programme that they have
  put in place for our country,
  South Africa.</seg>
```

When multi-paragraph foreign utterances are encountered, they are modelled one paragraph by one.

What is unusual is that meta-comments related to time expiration, immediately following the longer foreign language fragment, can be also rendered in a language other than English, e.g. "Kwaphela isikhathi." ("Time expired" in Zulu). Such cases were modelled in a standard way, with original Zulu text kept for reference:

```
<vocal type="interruption">
  <desc xml:lang="zu">Kwaphela
    isikhathi.</desc>
  <desc xml:lang="en">Time
    expired.</desc>
</vocal>
```

The time expiration note is put after both fragments, in the original language and English.

8 Speaker metadata

Following the ParlaMint model, the information of the speakers (MPs and guests) was stored in a separate file (`ParlaMint-ZA-listOrg.xml`), listing organisations involved in the process. In our experiments only two organisations were involved: the South African Government (represented by the Minister of Basic Education and her deputy) and the National Assembly Chamber of the 27th South African Parliament:

```
<org xml:id="government.ZA"
   role="government">
   <orgName xml:lang="en"
     full="yes">South African
     Government</orgName>
</org>
<org xml:id="parliament"
     role="parliament" ana="#epc">
```

```
    <orgName full="yes" xml:lang="en">
      National Assembly Chamber
    </orgName>
    <listEvent>
       <event xml:id="epc.27"
         from="2019-05-22">
         <label xml:lang="en">27th
           South African Parliament
         </label>
       </event>
    </listEvent>
</org>
```

The information about individual speakers were stored in `ParlaMint-ZA-listPerson.xml` file. There were 24 distinctive strings identifying speakers in the sample session. One of them was described as "An HON MEMBER" (original spelling) which corresponded to an unidentified speaker and his speech was encoded as a vocal message of some regular MP:

```
<vocal type="clarification"
  ana="#regular">
  <desc>It is a point of
    debate.</desc>
</vocal>
```

One name (Patamedi Ronald Moroatshehla) was twice misspelled as Moroatsehla so eventually, 22 speakers were encoded in the person file.

The names and roles of the speakers used in the transcript were converted into XML IDs by replacing spaces with underscores, deleting courtesy titles (Mr, Ms, Mrs, Dr etc.) and removing brackets to maintain compliance with `xml:id` type model, e.g. `The HOUSE CHAIRPERSON (Mr M L D Ntombela)` becomes `The_HOUSE_CHAIRPERSON_M_L_D_Ntombela`.

The surnames and initials of the first names of the MPs were then used to match against relevant web pages on the parliament website (Parliament of The Republic of South Africa 2023*b*). Even though various type of information was available on the MPs' websites such as their contact and social media details, political party and committee membership, parliament membership and committee membership history, political leadership background, education, interests and ambitions. Marital status could also be retrieved; while only "Ms" was used in the transcript, the website contains "Ms" and "Mrs" designations. They are not consistent for all MPs — cf. e.g. the information available for the chairperson Madala Louis David Ntombela (just basic data) and Nomsa Innocencia Tarabella Marchesi (very detailed description) so they were not included in the current experiment.

Eventually, the record of each speaker took the form:

```
<person xml:id="P_R_MOROATSHEHLA">
   <persName>
      <surname>Moroatshehla</surname>
      <forename>Patamedi
        Ronald</forename>
   </persName>
   <sex value="M"/>
   <idno type="URI"
     subtype="parliamentary-website">
       https://www.parliament.gov.za/
         person-details/244</idno>
   <affiliation ana="#epc.27"
      from="2019-05-22"
      ref="#parliament"
      role="member"/>
</person>
```

For the three MPs (Semakaleng Patricia Kopane, Désirée van der Walt and Lulama Maxwell Ntshayisa) who did not have their official sites on the website of the parliament, their information was retrieved from Wikipedia and their Wikipedia pages were linked instead of the parliament web pages. For Lulama Ntshayisa the reason was most likely his death of COVID in 2021 which was also noted in the person record.

When only speaker roles were given in the text instead of person names (the Minister of Basic Education and the Deputy Minister of Basic Education), actual names had to be manually retrieved from another source (Wikipedia). Since

they also had their records present at the parliament website, links to their pages were also included. When both the role and actual name were included, e.g. `The HOUSE CHAIRPERSON (Mr M L D Ntombela)`, both expressions were used.

The utterance information was additionally modified based on speaker metadata by setting the speaker type to:

- `chair` for all utterances marked as `#The_HOUSE_CHAIRPERSON_Mr_M_L_D_Ntombela`
- `guest` for all utterances of the Minister of Basic Education and her deputy
- `regular` was kept for all other utterances.

9 Summary and conclusions

The encoded session comprised of 95 utterances. 36 of them were by the chairperson, 16 by one of the MPs (Thlologelo Malatji), then we had two speakers with 7, one with 5 utterances, two with 3 and, three with 2 and twelve with only one utterance. Table 2 shows the distribution of individual constructs.

The proposed conversion process was intended only to illustrate how various decisions could be taken in the full-scale project of encoding South African parliamentary data in the ParlaMint schema. It by no means offers a final and definitive decision-making model and will have to be adjusted in the future. We are aware that it might not cover all phenomena encountered in real-world data since the encoding experiment covered only one session, one house and one main language. Still, we believe it can pave the way towards widening the European ParlaMint data with a new continent.

Acknowledgements

The work was supported by the 2014-2020 Smart Development Operational Programme, Priority IV: Increasing the scientific and research potential, Measure 4.2: Development of modern research infrastructure of the science sector, No. POIR.04.02.00-00C002/19, "CLARIN — Common Language Resources and Technology Infrastructure" and the Polish Ministry of Science and Higher Education, grant agreement 2022/WK/09.

References

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Fišer, D., ... & Kryvenko, A. (2023*a*), 'Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0'. Slovenian language resource repository CLARIN.SI.
URL: *http://hdl.handle.net/11356/1488*

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Fišer, D., ... & Kryvenko, A. (2023*b*), 'Multilingual comparable corpora of parliamentary debates ParlaMint 3.0'. Slovenian language resource repository CLARIN.SI.
URL: *http://hdl.handle.net/11356/1486*

Erjavec, T., Kopp, M. & Pančur, A. (2023), 'The structure and encoding of ParlaMint corpora'.
URL: *https://clarin-eric.github.io/ParlaMint/*

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M. & Fišer, D. (2023), 'The ParlaMint corpora of parliamentary proceedings', *Language Resources and Evaluation* **58**, 415–448.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014), 'The Sketch Engine: ten years on', *Lexicography* **1**, 7–36.

Kuzman, T., Ljubešić, N., Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Fišer, D. . & Kryvenko, A. (2023), 'Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana

Table 2: Statistics of ParlaMint elements in the encoded session

Element	Type / Reason	Count
note	time	2
	debate	2
	quote	2
kinesic	applause	19
	laughter	2
vocal	clarification	1
	interruption	8
	exclamation	35
gap	inaudible	1
	foreign	12

3.0'. Slovenian language resource repository CLARIN.SI.
URL: *http://hdl.handle.net/11356/1810*

Machálek, T. (2020), KonText: Advanced and flexible corpus query interface, *in* 'Proceedings of the 12th Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 7003–7008.
URL: *https://aclanthology.org/2020.lrec-1.865/*

Marneffe, de, M.-C., Manning, C. D., Nivre, J. & Zeman, D. (2021), 'Universal Dependencies', *Computational Linguistics* **47**(2), 255–308.
URL: *https://doi.org/10.1162/coli_a_00402*

Parliament of The Republic of South Africa (2023*a*), 'Proceedings of mini plenary session, 16 July 2019. National Assembly, Appropriation Bill'.
URL: *https://www.pa.org.za/hansard/2019/july/16/proceedings-of-mini-plenary-session-national-ass-2/appropriation-bill*

Parliament of The Republic of South Africa (2023*b*), 'Website of the Parliament of The Republic of South Africa: All Members'.
URL: *https://www.parliament.gov.za/group-details*

Parliament of The Republic of South Africa (2023*c*), 'Website of the Parliament of The Republic of South Africa: Disclaimer'.
URL: *https://www.parliament.gov.za/disclaimer*

Parliament of The Republic of South Africa (2023*d*), 'Website of the Parliament of The Republic of South Africa: Handard Papers'.
URL: *https://www.parliament.gov.za/hansard-papers*

Parliament of The Republic of South Africa (2023*e*), 'Website of the Parliament of The Republic of South Africa: Latest Hansards'.
URL: *https://www.parliament.gov.za/hansard*

Parliamentary Monitoring Group (2023), 'Hansard: Na: Unrevised Hansard (EPC), House: National Assembly, Date of meeting: 16 Jul 2019'.
URL: *https://pmg.org.za/hansard/28893/*

TEI Consortium, ed. (2017), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.
URL: *http://www.tei-c.org/Guidelines/P5/*

Tiedemann, J. & Thottingal, S. (2020), OPUS-MT – building open translation services for the world, *in* 'Proceedings of the 22nd Annual Conference of the European Association for Machine Translation', Lisboa, Portugal, pp. 479–480.
URL: *https://aclanthology.org/2020.eamt-1.61/*