The HOME Project Parliamentary Activity Data Sets

Full-text and annotated corpora of parliamentary activity in the legislatures of Ghana and South Africa.

YANI KARTALIS
Center for Social Data Science (SODAS), University of Copenhagen & Iscte – Instituto Universitário Lisboa
kartalisie@gmail.com

EDALINA SANCHES
Institute of Social Sciences, University of Lisbon
ersanches@ics.ulisboa.pt

JOÃO CONDUTO
Institute of Social Sciences, University of Lisbon
jpgarridoconduto@gmail.com

8 October, 2024

HOME – How Members of Parliament in Africa Represent their Constituencies Portuguese Foundation for Science and Technology Ref. PTDC/CPO/4796/2020
(https://doi.org/10.54499/PTDC/CPO-CPO/4796/2020)

Data Collection/Sources

The HOME Project parliamentary activity datasets cover the Ghanaian Parliament *(Parliament of Ghana*), and the South African Parliament *(Parliament of the Republic of South Africa)*. There are three separate datasets regarding the use of parliamentary tools by members of parliament contained in this dataset:

- The dataset of parliamentary **speeches** from the **Ghanaian Parliament** contains a comprehensive collection of speeches delivered between 7 January 2005 and 30 July 2024 and contains 352280 total utterances from 1679 unique debate files.
- The dataset of parliamentary **speeches** from the **South African Parliament** contains a comprehensive collection of speeches delivered between 14 June 1999 and 9 May 2024 and contains 309716 total utterances from 3342 unique debate files.
- The dataset of parliamentary **questions** from the **South African Parliament** contains a comprehensive collection of questions delivered between 14 June 2006 and 10 May 2024 and contains 61975 total questions from 1139 unique debate files.

Data on all speeches in both Ghana and South Africa were drawn from official hansard transcript documents in the respective parliament websites. Data on all questions in South Africa, submitted to both parliamentary houses (National Assembly & National Council of Provinces) are available on the website of the South African parliament in the form of the official question paper documents. There are two types of question papers according to the rules: The Question Paper produced for a particular question day or an Internal Question Paper containing all current questions that still must be replied to. The documents contain additional relevant information on party affiliation, date of submission and minister to whom the question was directed. Unfortunately, questions before 2006 are not available at the time of writing this document, which means we were unable to include the 1st (1994-1999) and 2nd democratic legislatures (1999-2004) and part of the 3rd legislature (2004-2009). The remaining data is complete, covering the 4th (2009-2014), the 5th (2014-2019), and the 6th (2019-2024) democratic parliaments.

For each of these parliaments and datasets the majority of the project's efforts went into correctly extracting speeches from the various document types (mostly .docx, .doc, or .pdf) and also correctly identifying MP's and assigning party and gender values.

Table 1, provides information with links to the sources used to get the raw data on plenary sessions. These online sources were scraped automatically using customized Python scripts adapting to the structures and formats of the respective database. For the cases where there were no machine-readable texts, an extra step of using Optical Character Recognition (OCR) software had to be applied to extract the textual information. We relied on Finereader 14 by ABBYY.

Dataset	Time Frame	Legislatures/Periods	Total Observations	URL	Collection Date
Ghana Speeches	2005 - 2024	4, 5, 6, 7, 8	352280	https://www.parliament.gh/docs?type=HS	29/09/2024
SA Speeches	1999 - 2024	2, 3, 4, 5, 6	309716	https://www.parliament.gov.za/hansard-papers	10/09/2024
SA Questions	2006 - 2024	3, 4, 5, 6	61975	https://www.parliament.gov.za/question-papers	10/09/2024

Table 1: Datasets Overview.

Variables

All three corpora have an identical column/variable structure with nine (9) variables.

The variable **Speech/Question** is a string variable that contains the raw text of the speech/utterance (in the Speeches datasets) or question (in the Questions dataset) as given in the original plenary transcript or question paper.

The variable **Speaker** indicates the name or title of the individual delivering the speech or tabling/asking the question, as extracted from the parliamentary transcripts or question papers. This string is sometimes reduced to only the name of the speaker in case it was accompanied by a role which was extracted to the respective column as below.

The variable **Date** stores the date when the speech was delivered or question tabled/asked in a character vector that has the format: (YYYY-MM-DD).

The variable **Legislature** is a string variable that contains the legislative period/term during which the speech was given or question tabled/asked.

The variable **Role** is a string variable indicating whether the person delivering the speech or tabling/asking the question had some parliamentary role in the specific legislature. Roles were not available in all cases and across years. However, the variable will always contain the information needed to distinguish between a speaker of the legislature and the acting chairperson during that session for the speeches datasets. When the variable takes the value "Chair" it specifies speeches given by the acting chairperson of the respective parliament or session. These speeches almost always serve for debate organisation only, no information on parliamentary group or even speaker name is normally provided and it is often useful to exclude them from analyses. In the case of the questions dataset, this variable is not available.

The variable **Gender** is a string vector that specifies the gender of the speaker.

The variable **Party** contains a string vector with the parliamentary group of the speaker as provided in the original transcript or as matched by external, time-stamped MP lists. Not always available.

The variable **HOME_Unique_id** is a string vector that specifies a unique identifier for each speaker, intended to allow matching with other HOME project datasets (namely the candidates, bios, and questions datasets).

The variable **Speech_len/Question_len** is a numeric variable that contains the overall length of the speech or question text calculated simply as a character count.

The variable **House** is a string variable only for the South African dataset, that contains information on whether the utterance was delivered in the National Assembly (NaAss) or the National

Council of Provinces (NCOP), Extended Public Committee (EPC - an old assembly chamber), and finally Joint Sessions (JOINT)

Additionally, for the Questions dataset **only**, the following variables are available: **Question-type** which contains information on the question type based on the type of document it was extracted from (read above), **Exec** which contains information on which ministry the question is directed to , **File_Name** which contains information on the name of the file it was extracted from, and finally, **Qnumber** which contains information of the question's internal reference number. All these could be used for future reference and correspondence with other datasets or documents from the South African Parliament.