The UN Security Council debates 1992-2023

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde van Meegdenburg, Antonio Pires

January 2025

1 Description

This dataset contains meeting protocols from the United Nations Security Council meetings between January 1992 and December 2023.

The protocols were downloaded as PDF documents from `http://www.un.org/en/sc/meetings/` and converted to text using OCR. In total, the corpus contains 106,302 speeches extracted from 6,233 meeting protocols of the United Nations Security Council. For a detailed description of the creation of the dataset as well as some basic statistics, please see our paper "The UN Security Council debates 1992-2023" (available on arxiv.org: `https://arxiv.org/abs/1906.10969`).

tokens	types	sentences
66392703	27582017	2341293

2 Citation

When using our dataset, please use the citation indicated on the Dataverse's website.

> Schoenfeld, Mirco; Eckhard, Steffen; Patz, Ronny; Meegdenburg, Hilde van; Pires, Antonio; 2019, "The UN Security Council Debates", https://doi.org/10.7910/DVN/KGVSYH, Harvard Dataverse, V2

3 Files

speeches.tar This is an archive of text files with single speeches in them. Every text file's filename contains an identifier for the meeting, a year, and an ordering number for every speech of the corresponding meeting. The first part of the filename equals the filename of the corresponding meeting in `txt.tar`.

speaker.tsv This is a tabulator-separated text file containing metadata for every speech. See section 5 for an explanation of fields.

meta.tsv This is a tabulator-separated text file containing metadata for every Security Council meeting. See section 5 for an explanation of fields.

docs.RData This is an `R`-object containing raw texts of all speeches. See section 5 for an explanation of fields. The object's structure allows for an easy conversion to a `quanteda`-Corpus object for further processing, for example.

docs_meta.RData This is an `R`-object containing metadata of all speeches and Security Council meetings. The contained dataframes are a combination of `speaker.tsv` and `meta.tsv`. See section 5 for an explanation of fields.

4 Sources

This dataset is compiled semi-automatically from meeting protocols downloaded from `http://www.un.org/en/sc/meetings/`. Please see our paper "The UN Security Council debates 1992-2023" (available on arxiv.org: `https://arxiv.org/abs/1906.10969`) for a detailed description of the compilation-procedure.

Please contact mirco.schoenfeld@tum.de for any inaccuracies you may encounter.

5 Explanation of variables

year – The year of the speech

month – The month of the speech

day – The day of the speech

date – The date of the speech in the form `DD Month YYYY`.

spv – The "S/PV"-meeting record identifier. The correct identifier is constructed by S/PV.{entry}. If the entry contains 'Resumption', the correct identifier is constructed by encapsulating the 'Resumption'-part as well as the following number in parentheses.

document_symbol – The full document symbol where the speech is recorded.

speech_number – An ongoing number numbering the speeches of one meeting.

speaker – The name of the speaker.

country_org – The country or organization the speaker represents during the speech. This is indicated either in the participants list of a meeting protocol or the beginning of the speech if the speaker is not listed as part of the participants list. If this field contains "UN", the speakers affiliation was checked manually and the following column contains more specific information.

role_in_un – The speaker's job in the UN, e.g. "Secretary General".

participanttype – Expresses the role of the speaker. This field has one of the following three entries:

- `The President`: The speaker was the meeting's president
- `Mentioned`: The speaker is listed in the participants list of the covering page of the protocol.

- `Guest`: The speaker is not listed in the participants list but rather invited to take a seat at the table during the meeting.

content – **only in R dataset** – The actual text of the speech

filename – The filename of the speech given in `speeches.tar`.

basename – The basename of the meeting protocol in `txt.tar` without the file extension (`.txt`).

topic – The topic of the meeting indicated in the agenda section of the PV.

agenda_item1 – The first layer of agenda item label according to the Repertoire of the Practice of the UNSC, which includes six categories: Africa, Americas, Asia, Europe, Middle East, and Thematic Items.

agenda_item2 – The second layer of agenda item label according to the Repertoire of the Practice of the UNSC.

agenda_item3 – The third layer of agenda item label according to the Repertoire of the Practice of the UNSC.

agenda_item4 – The fourth layer of agenda item label according to the Repertoire of the Practice of the UNSC.

agenda_item_manual – Manual label for agenda item (therefore non-official) based on the Repertoire of the Practice of the UNSC.

types – The number of unique tokens in the speech.

tokens – The number of tokens in the speech.

sentences – The number of sentences in the speech.

num_speeches – The number of speeches held during the corresponding meeting.

speech_format – Whether the speech was delivered in person or via video-teleconference (VTC).