The UN Security Council debates 1995-2020

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

July 2021

1 Description

This dataset contains meeting protocols from the United Nations Security Council meetings between January 1995 and December 2020.

The protocols were downloaded as PDF documents from `https://research.un.org/en/docs/sc/quick/meetings/` and converted to text using OCR. In total, the corpus contains 82,165 speeches extracted from 5,748 meeting protocols of the United Nations Security Council. For a detailed description of the creation of the dataset, please see our paper "The UN Security Council debates 1995-2017" (available on arxiv.org: `https://arxiv.org/abs/1906.10969`).

tokens	types	sentences
52311385	21480773	1834979

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, 2021, "The UN Security Council Debates", https://doi.org/10.7910/DVN/KGVSYH, Harvard Dataverse, V5

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 `https://research.un.org/en/docs/sc/quick/meetings/`. Please see our paper "The UN Security Council debates 1995-2017" (available on arxiv.org: `https://arxiv.org/abs/1906.10969`) for a detailed description of the compilation-procedure.

Please contact mirco.schoenfeld@uni-bayreuth.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.

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

speaker   – The name of the speaker.

country   – The country or affiliation 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.

national_capacities   – **only in R dataset** – a list of speeches in which the President spoke in capacity of her nation. Together with the speeches listed in *un_capacities* and *prst_speeches_filenames*, these speeches are all non-procedural speeches given by the President. Removing every other presidential speech not in one of the three lists means to filter all procedural speeches ("I now give the floor [...]").

un_capacities   – **only in R dataset** – a list of speeches in which the President spoke in capacity of the UN, for example if she delivered a report of a Security Council Committee. Together with the speeches listed in *national_capacities* and *prst_speeches_filenames*, these speeches are all non-procedural speeches given by the President. Removing every other presidential speech not in one of the three lists means to filter all procedural speeches ("I now give the floor [...]").

prst_speeches_filenames   – **only in R dataset** – a list of speeches in which the President spoke in capacity of the Security Council. Together with the speeches listed in *un_capacities* and *national_capacities*, these speeches are all non-procedural speeches given by the President. Removing every other presidential speech not in one of the three lists means to filter all procedural speeches ("I now give the floor [...]").

agenda_item1   – the broadest topic labels, based on "Agenda Items: Overview", on the Repertoire Website[1]: Africa, Americas, Asia, Europe, Middle East, and Thematic Items

agenda_item2   – the second broadest topic labels, based on the titles available in each section of "Agenda Items: Overview". It's constituted mainly by countries' names, except in the case of Thematic Items;

topic2   – the third broadest topic labels, based on the subtitles available in each section of "Agenda Items: Overview". They're also present in the Repertoire of the Practice of the Security Council Supplements after a number[2];

subtopic   – the fourth broadest topic labels, present in the Repertoire of the Practice of the Security Council Supplements. These labels usually come after a letter[3] or at the end of a chapter[4].

agenda_item3   – the more specific topic labels. These are not official labels present in the Repertoire but were created according to information present in *subtopic* and *agenda_item2*, to describe more clearly the situation under discussion, focusing on countries and world regions. The Thematic Issues were kept as "Thematic" in this variable. Clear and specific information about Thematic Issues can be viewed in *agenda_item2*. **Observations:** "The situation in the occupied Arab territories" was renamed to "Israel/Palestine". Meeting 4862 was renamed to "Israel/Palestine". "Security Council resolution 1595 (2005)" and "Security Council resolution 1559 (2004)" were renamed to "Lebanon". "UNIFIL" was renamed to "Israel/Lebanon". "UNDOF" was renamed to "Israel/Syria". Meeting 4863 was renamed to "Israel/Palestine".

---

[1]See: `https://www.un.org/securitycouncil/content/repertoire/structure`, last visited 16 July 2021

[2]See: `https://www.un.org/en/sc/repertoire/2016-2017/Part_I/2016-2017_Part_I.pdf#page=135` p. 135 ("24. The situation in the Middle East").

[3]See: `https://www.un.org/en/sc/repertoire/96-99/Chapter8/MiddleEast/96_99_8_MiddleEast_30_MiddleEast.pdf` p. 927. ("B. United Nations Interim Force in Lebanon and development in the Israel-Lebanon sector").

[4]See: `https://www.un.org/en/sc/repertoire/2016-2017/Part_I/2016-2017_Part_I.pdf#page=135` p. 140. ("the situation in the Middle East - Syrian Arab Republic").

decision – this variable indicates if a PRST was adopted and if the veto was used. It needs to be improved with more detailed information.

text – **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 protocol.

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.

outcome – A link to the outcome of the meeting as indicated on the website of the UN (`https://research.un.org/en/docs/sc/quick/meetings/`).

pressrelease – A link to the press release of the meeting as indicated on the website of the UN (`https://research.un.org/en/docs/sc/quick/meetings/`).