**United National General Assembly Oral Interventions Dataset**

Spyros Blavoukos and Ioannis Galariotis

This document provides supplementary information for the paper entitled “Oral Interventions in the UN General Assembly: Accounting for the Variance in EU Member-States’ Engagement” concerning the dataset we have used in our analysis and, more specifically, it elaborates on the following items:

1. The data sources and collection guide;
2. Descriptive statistics of the datasets;
3. Collection items and rules for oral interventions.

All files and analysis can be found in the following GITHUB link:

[**https://github.com/igalario/EU\_oral\_interventions**](https://github.com/igalario/EU_oral_interventions)

1. **Data sources and Collection Guide**

**1.1 United Nations General Assembly Plenary Meetings**

We have collected all the official documents that are available from the 53rd to the 72nd plenary session. To download these documents, one could go to the following weblink:

<https://research.un.org/en/docs/ga/quick/regular/77>.

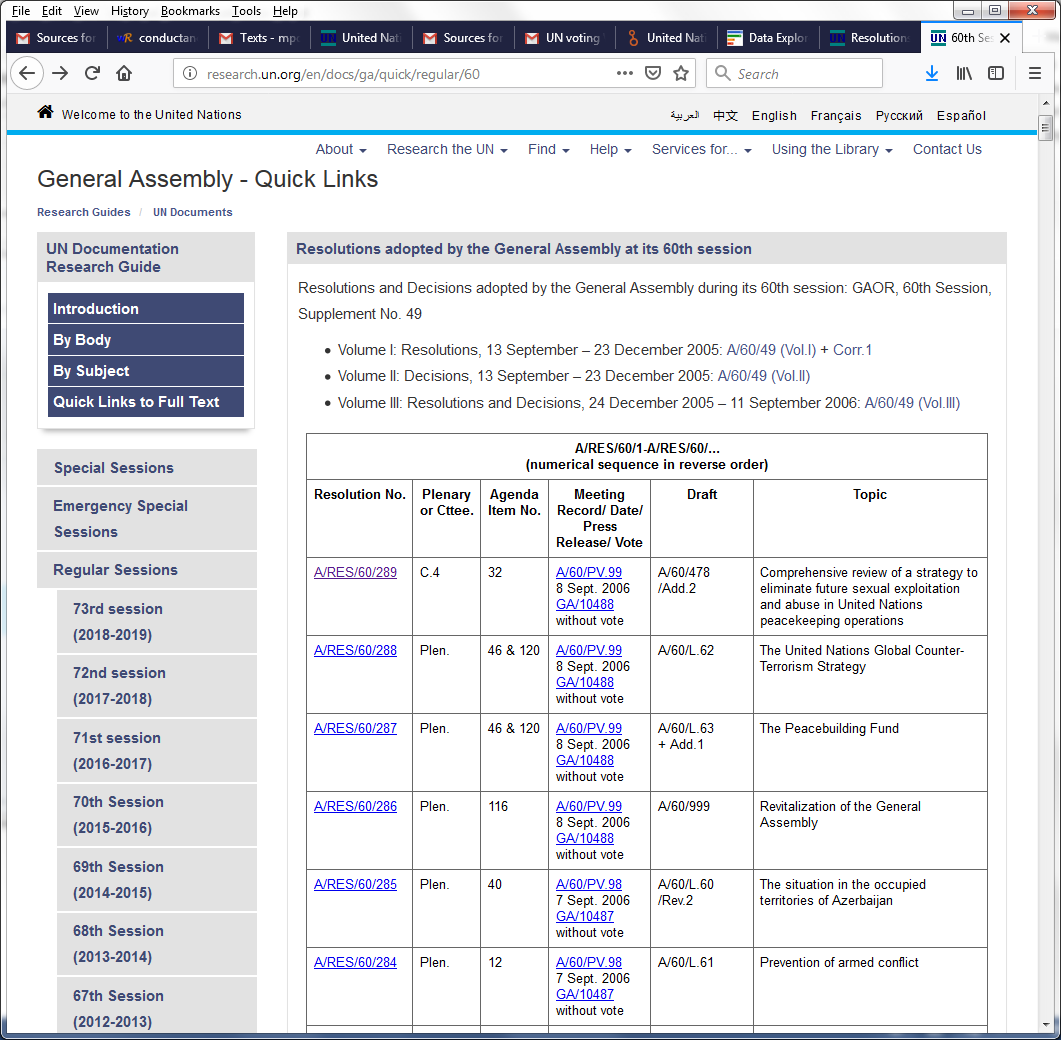
We have downloaded all the meetings with the following format for each plenary session:

[A/Number of the Plenary Session/Number of PV plenary meeting]

For instance, regarding the 60th session, one document is the following: A/60/PV.99 and the link for this document is the following:

[**https://documents-dds-ny.un.org/doc/UNDOC/GEN/N06/493/43/PDF/N0649343.pdf?OpenElement**](https://documents-dds-ny.un.org/doc/UNDOC/GEN/N06/493/43/PDF/N0649343.pdf?OpenElement)**.**

The following screenshot presents an example for the 60th plenary session and the documents we have downloaded.



* 1. **United Nations General Assembly Main Committees Meetings**

The United Nations General Assembly consists of six Main Committees:

1. Disarmament and International Security (First Committee)
2. Economic and Financial Committee (Second Committee)
3. Social, Humanitarian & Cultural Issues (Third Committee)
4. Special Political and Decolonization (Fourth Committee)
5. Administrative and Budgetary Committee (Fifth Committee)
6. Legal (Sixth Committee)

We have gathered documents from the 1st, 2nd, 3rd, 4th and 5th Main Committees of the United Nations General Assembly. The 6th Main Committee is responsible for the legal questions in the United Nations General Assembly, and, because it is out of the scope of our analysis, we have excluded it from our dataset collection.

The United Nations has a digital library where someone can search, trace and download a wealth of text information. The link of the United Nations Digital Library is the following:

[**https://digitallibrary.un.org/collection/General%20Assembly?ln=en**](https://digitallibrary.un.org/collection/General%20Assembly?ln=en)

We have downloaded all the meetings with the following format for each Main Committee session:

[A/Main Committee/Number of the Plenary Session/Number of PV plenary meeting]

For instance, regarding the 72nd session of the 4th Main Committee, one document is the following: A/C.1/72/PV.21 and the link for this document is the following:

<https://digitallibrary.un.org/record/1654118?ln=en>

1. **Descriptive statistics of the datasets**

We have gathered **4835** pdf files in total following the data collection guide described in the previous section. In particular, we have gathered **842** pdf files [sessions 53(1998) – 72(2017)] from the Plenary and **3993** pdf files from the five Main Committees, i.e. Main Committee 1 (C1), Main Committee 2 (C2), Main Committee 3 (C3), Main Committee 4 (C4), and Main Committee 5 (C5).

In the following table,

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Number of pdf files per session** | | | | | | |
| **No of Session** | **Plenary** | **C1** | **C2** | **C3** | **C4** | **C5** |
| 53 | 49 | 24 | 43 | 54 | 25 | 70 |
| 54 | 45 | 25 | 53 | 56 | 26 | 75 |
| 55 | 51 | 28 | 42 | 58 | 29 | 71 |
| 56 | 39 | 22 | 41 | 64 | 24 | 62 |
| 57 | 33 | 25 | 46 | 64 | 26 | 57 |
| 58 | 33 | 24 | 41 | 64 | 26 | 53 |
| 59 | 38 | 27 | 43 | 57 | 29 | 60 |
| 60 | 36 | 21 | 60 | 49 | 29 | 68 |
| 61 | 41 | 24 | 36 | 56 | 30 | 58 |
| 62 | 33 | 23 | 37 | 56 | 26 | 52 |
| 63 | 46 | 25 | 33 | 51 | 28 | 55 |
| 64 | 48 | 24 | 43 | 49 | 28 | 37 |
| 65 | 42 | 26 | 34 | 53 | 28 | 43 |
| 66 | 44 | 24 | 41 | 51 | 25 | 39 |
| 67 | 46 | 24 | 39 | 46 | 24 | 37 |
| 68 | 42 | 24 | 42 | 53 | 27 | 46 |
| 69 | 47 | 24 | 39 | 56 | 27 | 45 |
| 70 | 42 | 24 | 38 | 57 | 27 | 38 |
| 71 | 43 | 25 | 30 | 58 | 25 | 39 |
| 72 | 44 | 25 | 29 | 54 | 30 | 45 |
| Total | 842 | 488 | 810 | 1106 | 539 | 1050 |

The obtained pdf files had been processed and transformed to txt files (machine readable format).

All the raw data (pdf and txt files) are available in the GITHUB link in the first page of this document.

1. **Collection items and rules for oral interventions**

**3.1 GENERAL GRAMMAR UNIT**

Given as input a preprocessed pdf file either from the Plenary or a Main Committee session, the GENERAL GRAMMAR UNIT has been designed to identify the following types of information from the raw text:

* DATE
* SESSION
* GA BODY
* MEETING
* PV

In addition, two GRAMMAR UNITs have been designed and implemented, one for each main United Nations body, i.e. Plenary or Main Committees:

* ComGeneral (consisting of 7 modules)
* PlenGeneral (consisting of 7 modules)

Each GRAMMAR UNIT, i.e. ComGeneral of PlenGeneral, comprises of the following modules containing rules for the detection of each type of information:

**Module 1: Date**

**Module 2: meeting**

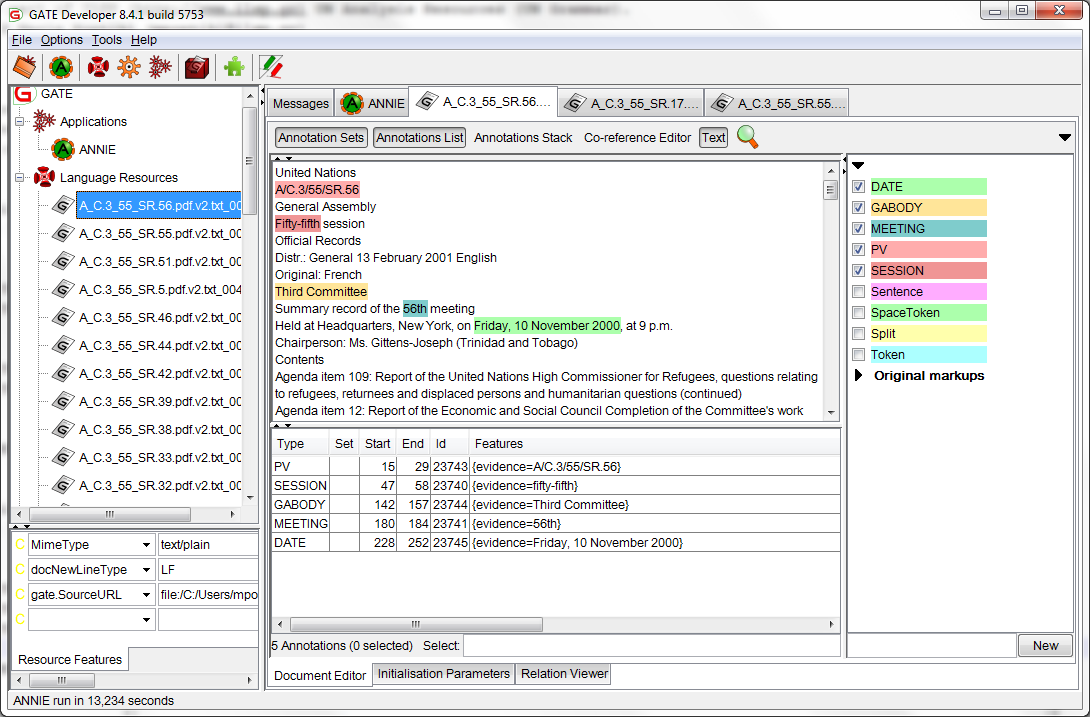
**Module 3: session**

**Module 4: GA body**

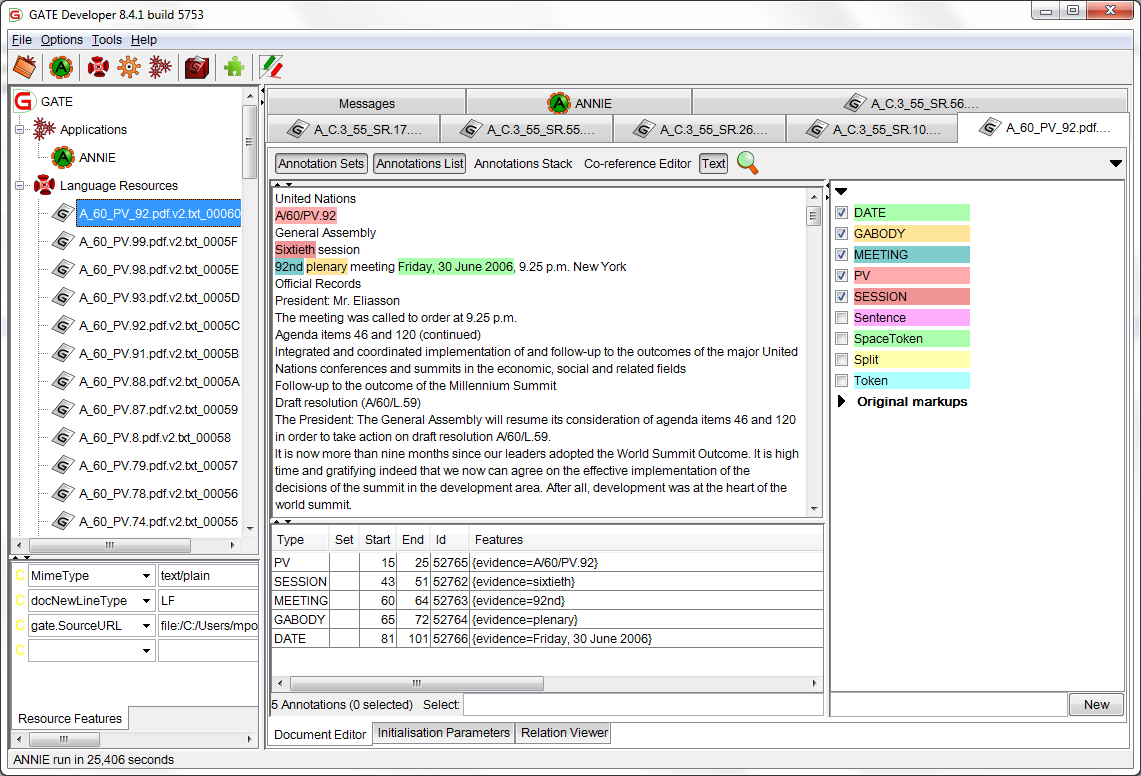
**Module 5: PV modules**

In addition, they include a “Multitoken” module (**Module 6**) for extracting the linguistic evidence of multitoken information (PV, Date and in the case of committees also GA body) and a RejectTemp module (**Module 7**) that cleans temporary labels. For each type of the identified info the GRAMMAR UNITs return also the respective linguistic evidence from the text as illustrated below in Figures 1 and 2:

**Figure 1: Example of ComGeneral Grammar output**



**Figure 2: Example of PlenGeneral Grammar output**



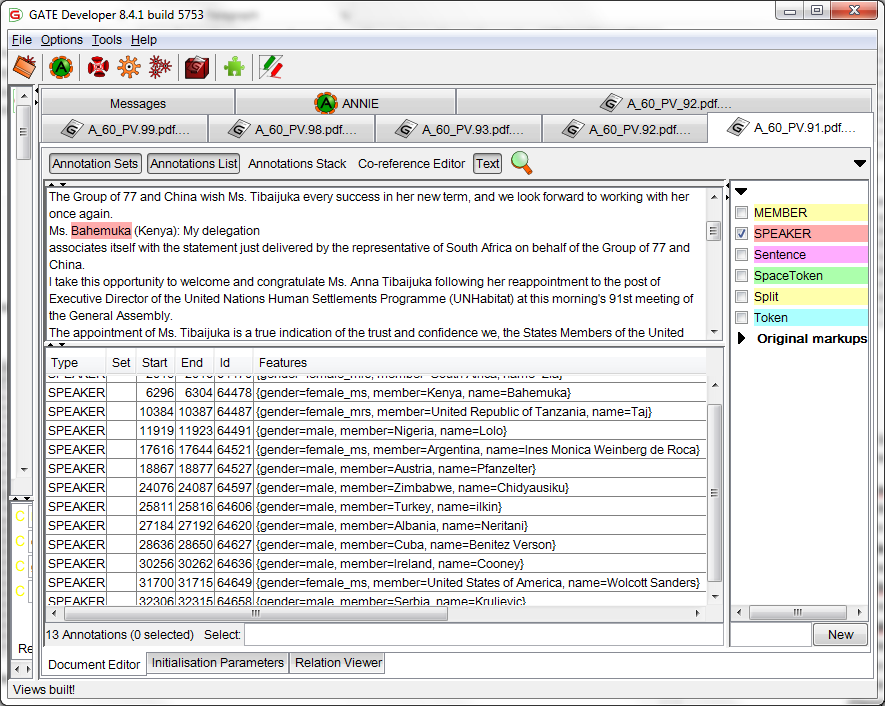
**3.2 SPEAKER GRAMMAT UNIT**

The SPEAKER GRAMMAR UNIT detects speakers representing UN member states and expressing positions (i.e. not everyone who speaks during a meeting such as the president/chairperson of the meeting or observers from other countries – more examples are provided below at SpeakReject module). For each identified speaker, the SPEAKER GRAMMAR UNIT returns the following types of information:

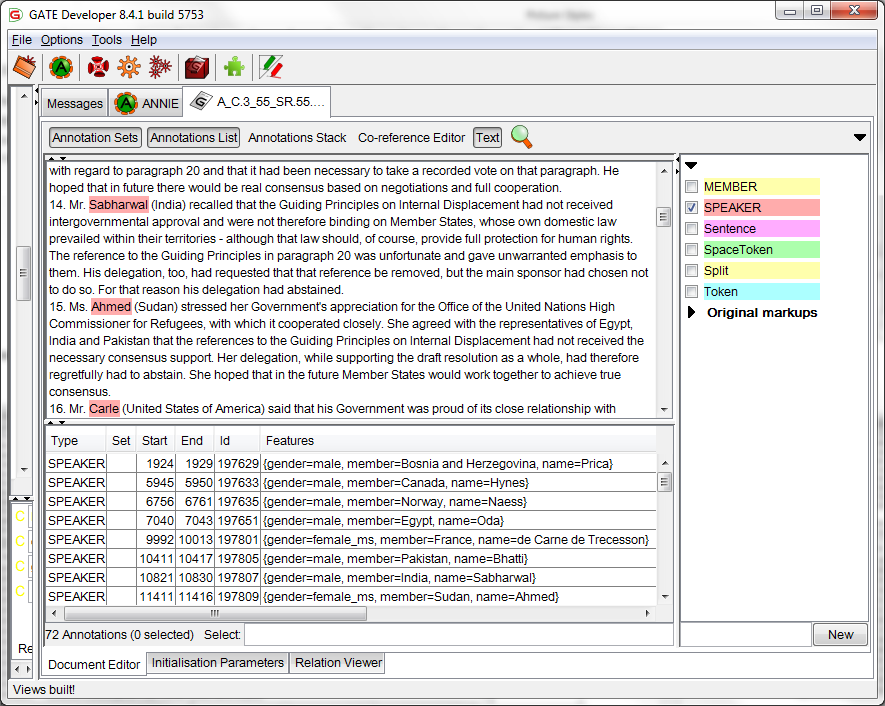
* ***Name***: the name of the Speaker e.g. *Gert Auvaart, Gonsalves, Pereira Neves, De Aguiar Patriota, Ja'afari, B0jer, Alarcdn de Quesada, Han Sung Il, SimonoviC, Samhan Al-Nuaimi, Ja Song Nam*
* ***Gender***: *male* or *female*. In the case of female speakers, the title (mrs or ms) is also included.
* ***Member***: the member state the speaker represents e.g. *Denmark, Brazil, United Arab Emirates, Democratic People's Republic of Korea.*

Examples of the SPEAKER GRAMMAR UNIT output from the Plenary and the Main Committee meeting minutes are illustrated below in Figures 3 and 4:

**Figure 3: Example of SPEAKER GRAMMAR UNIT output from Plenary’s meeting minutes**



**Figure 4: Example of SPEAKER GRAMMAR UNIT output from Main Committee’s meeting minutes**



The SPEAKER GRAMMER UNIT is common for both the Plenary and the Main Committees and it comprises of the following 5 modules:

* **UN\_members module**: Set of patterns for detecting lexicalizations of the names of member states of the United Nations over time. Note that the European Union is identified separately as a member-entity, when it is represented.
* **Speaker module**: Set of patterns that generalize rules for detecting speakers and the respective annotation features (name, gender, member).
* **Multitoken module**: Set of patterns for extracting the linguistic evidence of multitoken information.
* **SpeakReject module**: Set of patterns that generalize rules for rejecting candidate speakers when identified in particular contexts e.g.

*In the absence of the President, Mr. Sevilla Borja (Ecuador) took the Chair*.

*At the invitation of the President, Ms. Moreno Gonzalez (Ecuador), Mrs. Kanchaveli (Georgia), Ms. O'Brien (Ireland), Mr. Francis (Lebanon), Mr. Perez Ayestaran (Bolivarian Republic of Venezuela) and Mr. Silwamba (Zambia) acted as tellers.*

*The President: Mr. Ashe was a friend and colleague of many here in the Hall.*

* **TempReject module**: 2 patterns that clean previously assigned temporary labels.