The UNSC-Graph: An Extensible Knowledge Graph for the UNSC Corpus

Stian Rødven-Eide
Språkbanken Text
University of Gothenburg
stian.rodven.eide@svenska.gu.se

Karolina Zaczynska
Applied Computational Linguistics
University of Potsdam
zaczynska@uni-potsdam.de

Antonio Pires
Political Science and International Relations
Federal University of Pernambuco
antonio.hps26@gmail.com

Ronny Patz
Faculty of Human Sciences
University of Potsdam
ronny.patz@uni-potsdam.de

Manfred Stede
Applied Computational Linguistics
University of Potsdam
stede@uni-potsdam.de

Abstract

We introduce the UNSC-Graph, a knowledge graph for a corpus of debates of the United Nations Security Council (UNSC) during the period 1995-2020. The graph combines previously disconnected data sources including from the UNSC Repertoire, the UN Library, Wikidata, and from metadata extracted from the speeches themselves. Beyond existing metadata detailing debates' topics and participants, we also extended the graph to include all country mentions in a speech, geographical neighbours of countries mentioned, as well as sentiment scores. By linking the graph to Wikidata, we are able to include additional geopolitical information and extract various country name aliases to extend the coverage of country mentions beyond existing NER-based approaches. Studying mentions of Ukraine after 2014, we present a use case for the graph as a source for continuous analysis of international politics and geopolitical events discussed in the UNSC.

1 Introduction

Since 1946, the United Nations Security Council has been the most important global body for discussion and action pertaining to global security. Its regular meetings are among the most publicly visible in the UN (Schönfeld et al., 2019). Although many conversations among permanent and non-permanent UNSC members take place behind closed doors, the documentation of the Security Council's public meetings is an essential source for studying international conflicts and threats to peace and security.

In 2019, and updated in 2021, a corpus of complete meeting transcripts from 1995 to 2020 (originally to 2017), with cleaned text and a range of metadata, was released by Schoenfeld et al. (2021).[1] In this paper, we build on this corpus to present the UNSC-Graph, a knowledge graph built with Prolog, a programming language that is practical for building knowledge graphs and can be addressed via Python and R. The UNSC-Graph expands, enhances, and augments the UNSC corpus. As a result, a new range of research questions can be addressed in political science beyond what is possible with the current corpus and metadata. Most notable is the improvement of detecting country mentions including different variations of country names while having additional context such as whether this country is in the room, is a neighboring country, or has been a UNSC member in the past.

2 Related Work in PolSci and CompLing

Various research projects have been examining the existing UNSC corpus since its release, notably in political science and computational linguistics.

[1] Earlier meetings can only be examined as summaries, most of which are only publicly available as scans of typed text

Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences, pages 69–74

Rhetoric analyses include Medzihorsky et al. (2017), focusing on the debates surrounding the Syrian civil war, and Bakalova and Jüngling (2020), which compares US and Russian rhetoric. Scherzinger explores new methods of quantitative rhetorical analysis (2023b) and sentiment shifts around "R2P" (2023a).

Discourse and network analyses abound on UNSC topics ranging from Afghanistan debates (Eckhard et al., 2023) to climate change (Scartozzi, 2022), health issues (Voss et al., 2022), and Russia's invasion of Ukraine (Bendix, 2022). UNSC discourse is also discussed in (Campbell and Matanock, 2021) and (Badache et al., 2022).

Among the more linguistically oriented computational investigations relevant for this paper are a named entity extension by Glaser et al. (2022) and an analysis of country mentions by Ghawi and Pfeffer (2022). The former uses Wikidata for Named Entity Linking (NEL) in the UNSC corpus, including for recognising country mentions, but it builds on the more complex Resource Description Framework (RDF) and does not make use of the wider knowledge base presented in this paper. The latter paper constructed a dataset of country mentions within the UNSC debates and mapped them to ISO names via named entity recognition (NER). For our graph, we opted to use string matching instead of NER, and incorporated different aliases for country names provided by Wikidata to capture a wide range of mentions, as well as information about the country's membership role during the meeting. We were able to detect a total of 768,131 country mentions, a significant improvement over the 211,237 mentions in Ghawi and Pfeffer (2022).

Outside of the UNSC domain but in the domain of political debates, knowledge graphs have become more prevalent. The ParliamentSampo knowledge graph includes data about MPs, parliamentary speeches, and political organizations within the Parliament of Finland (Hyvönen et al., 2022). Tamper et al. (2022) enrich the knowledge graph of the plenary debates extracting named entities and topical keywords using NLP methods. LinkedEP is a Linked Open Data version of European Parliament's data (Van Aggelen et al., 2016). Linked data has also been used in the LinkedSaeima for the Latvian parliament (Bojārs et al., 2019). The Swedish PoliGraph (Rødven-Eide, 2019), a knowledge graph for Swedish parliamentary debates, utilises Prolog to simplify the quest for answers pertaining to debates and their respective metadata, which informs the use of Prolog for building the UNSC-Graph.

3 Creating the UNSC-Graph

The original English-language dataset from Schoenfeld et al. (2021) is available through the Harvard Dataverse as plain text and R-files.[2] In addition to the speeches themselves, metadata for both speeches and meetings provide information on meetings participants and speakers, time and topic(s) of each debate.

The corpus contains 82,165 speeches from 5,223 meetings between 1995 and 2020. As the speeches retained line breaks from the PDFs they were extracted from, we first recreated the corpus with reconstructed sentences and submitted this updated version to the authors for perusal in future releases. The sentence count is 1,685,801, which gives an average of 20.5 sentences per speech.

In a second step, all names of countries that participated in the UNSC debates during the period of the corpus were resolved and linked to their respective Wikidata-IDs. In order to detect mentions of countries that did not participate, we augmented this with a list of "`instance of country or sovereign state or state with limited recognition`" through Wikidata's SPARQL service. This list included several historical countries where the naming could be ambiguous to modern countries. We manually removed those for which that could be the case. Since some of the countries are referred to by different names in the UNSC corpus and in Wikidata, we kept two different sets of official names for each country, here labelled `wdlabel` and `unterm`. Wherever the UN data did not include an official term, we set it to the value of the former.

Third, we assembled a list of alternative strings that could reasonably be used to refer to a country – such as e.g. "Holland" for the Netherlands or "Burma" for Myanmar – by obtaining all aliases for each country from Wikidata, only filtering out those that were shorter than four characters.

For all country mentions, we used queries on the aliases obtained from Wikidata to search the text for strings referring to specific countries, and then resolving them to their corresponding Wikidata-ID. In contrast to the method used by Ghawi and

[2] `https://doi.org/10.7910/DVN/KGVSYH` (URLs were all last accessed on 2023-06-14.)

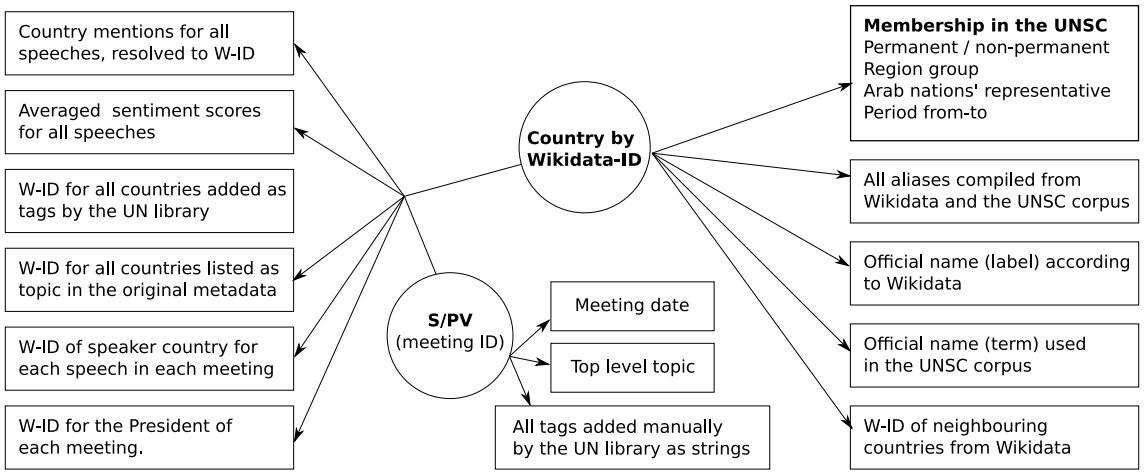

Figure 1: An illustration of the UNSC-Graph, detailing available data for research on the UNSC corpus.

Pfeffer (2022), we deemed NER as an unnecessary step in this context and were able to detect 768,131 country mentions. These were then subsequently incorporated into the graph as well. We opted to document the mentions on sentence level, which means that any given country is only counted as mentioned once per sentence. Along with the speaking country's and the mentioned country's respective Wikidata-ID, we stored the meeting and speech in which the mention took place, as well as the paragraph and sentence number.

Fourth, we added detailed membership information for all members of the UNSC from 1995 until today, as well as the geographical neighbours of each country, from Wikidata. This allows exploring the difference between mentions of neighbouring and other countries when UNSC participants debate a given conflict or topic, or to see whether non-permanent members argue differently from the Council's permanent members or non-members.

Fifth, using the existing metadata supplied with the corpus, we assembled lists of dates, topics, and presidency information for each meeting, as well as the Wikidata-ID of the country of the speaker for each speech. We also obtained a novel list of topic tags for each meeting that were manually crafted by the UN Library, which is responsible for UNSC document management. These tags are are based on the UNBIS taxonomy.[3] It provides a controlled vocabulary for describing UN documents, enabling semantic (topic-based) searching by identifying documents on the same concept, consistent with changing terminology. These tags complement the topic labels already available from the UNSC repertoire that are in the existing metadata. Each tag was coupled with the corresponding meeting identifier as a string, and for each that referred to a country, that country's Wikidata-ID was also linked to the meeting as a tag.

Sixth, we have added the average sentiment scores for each speech using the dictionary-based Lexicoder library, designed for sentiment analysis in political texts (Young and Soroka, 2012).

The assembled knowledge from a diverse set of sources was then stored as Prolog facts, using the SWI-Prolog implementation[4]. For a specialised knowledge graph such as the UNSC-Graph, we suggest that Prolog is a particularly good alternative to RDF-based solutions, as it enables rapid prototyping and lowers the threshold for both creating, storing, and using the graph. Prolog facts, especially in a project of small to medium size, can be stored as plain text files. As Prolog is both multi-directional and modular, with predicates supporting any number of arguments, we avoid the reliance on RDF triplets and offer a solution that easily can be extended and modified without any Prolog knowledge required. The modular nature of Prolog means that any research project wishing to utilise the UNSC-Graph easily can craft additional knowledge and rules to assist with complex queries. Outside of the Prolog environment itself, there are no requirements for running or querying the graph. The graph is Free Software[5], licensed under GPLv3[6] or later, and can be found at https://codeberg.org/Stian/UNSC-Graph.

[3] https://research.un.org/en/thesaurus

[4] https://www.swi-prolog.org/

[5] https://fsfe.org/freesoftware/freesoftware.en.html

[6] https://www.gnu.org/licenses/gpl-3.0.en.html

Country	Mentions/speech	Std. deviation	Mentions	Speeches	Avg. sentiment
Lithuania	2.37	10.21	467	197	-0.0017
Luxembourg	1.52	8.69	155	102	0.0153
Australia	1.35	10.73	216	160	0.0138
Russia	1.06	10.61	1201	1131	0.0076
Rwanda	0.93	6.95	129	138	0.0172
Latvia	0.83	2.47	20	24	0.0241
Jordan	0.69	5.90	159	232	0.0439
USA	0.61	11.83	597	984	0.0131
Chad	0.58	3.70	97	166	0.0219
Argentina	0.54	4.43	89	165	0.0261
Chile	0.50	5.14	107	215	0.0444
UK	0.47	9.42	461	987	0.0145
China	0.18	5.41	172	942	0.0609

Table 1: Excerpt from statistics of speeches that mention Ukraine in the UNSC debates after 1 March 2014.

As many researchers prefer to work with Python or R, we refer to the PySwip library[7] and the rolog[8] package for the respective programming languages. Python examples are included in the repository for the UNSC-Graph, as well as documentation of all available predicates. An illustration of the resulting graph can be seen in Figure 1.

4 Using the Graph: Mentions of Ukraine

We designed the UNSC-Graph with flexibility and modularity in mind. As a result, it can be used for a wide range of research questions. One particular area of application is the analysis of country mentions. In general, speakers in the Security Council represent countries and are usually also referred to as such by others. Since much of a debate deals with various countries' relation to each other or conflicts within or between countries, country mentions are both prevalent and meaningful. A quick count of string matches of aliases shows that as much as 34.36% of corpus sentences contain a country name. This number does not even include the many alternative ways of referring to a nation, such as the name of their capital.

The UNSC-Graph can easily tell us whether a country mentioned is taking part in the same meeting or not, whether it has already spoken or whether it will be speaking later in that meeting, as well as who currently is a member of the UNSC. We can e.g. see that Ukraine speaks in meetings on Crimea even when it is not a member. Furthermore, we can easily ascertain when a given country is listed as a topic for the meeting, officially as per the Security Council's own category (the original metadata of the corpus), or according to the UN Library's analysis (tags added for this paper). This is relevant because mentions of Ukraine in a UNSC meeting on Crimea may simply be a function of the meeting itself, while Russia mentioning Ukraine in a UNSC meeting on the general topic of terrorism may indicate its general focus on Ukraine.

Focusing on the mentions of Ukraine between the start of the current conflict with Russia in March 2014 and December of 2020, we find a total of 9,143 mentions. By complementing the query with Prolog rules, we then excluded speeches by the meeting president, as these usually consist only of formal announcements of topics and participants. This resulted in 8,382 mentions, of which 2,319 are by Ukraine themselves. For the remaining 6,063 mentions, we analysed the mentioning country and sorted these by the average mentions of Ukraine by those speakers.

In Table 1, we show the top 12 countries with the highest share of Ukraine mentions per speech, adding China as an important permanent member. The resulting table, using just some of the information from the UNSC-Graph, shows, for example, that Russia mentions Ukraine most frequently and in speeches that are generally much more negative than for any of the other countries shown here except for Lithuania. In contrast, China mentions Ukraine six times less frequently than Russia, and in speeches that are much more positive than the others, showing China's practice of avoiding finger pointing by not mentioning conflict parties directly.

[7] https://github.com/yuce/pyswip
[8] https://github.com/mgondan/rolog/

5 Future Work and Conclusion

This paper describes the design, generation and potential use cases of a new knowledge graph for the UNSC corpus, combining existing metadata with information from Wikidata and the UN Library, as well as several extensions by means of natural language processing methods. The graph facilitates content analysis using NLP methods, such as the vocabulary used by diplomats and countries in their speeches, the detection of conflict between speakers in the room, or the choices of speakers to focus on or ignore certain conflicts or conflict parties.

Considering that the UNSC-Graph uses Wikidata's identifiers for countries, extending the graph further – either through more information from Wikidata or through other linked open datasets – is a natural next step, depending on the directions of future research projects. Further work to extend the UNSC-Graph may include expanding more topic information from the UN Repertoire or any other knowledge base researchers in political science or NLP may want to add for their respective research questions.

6 Limitations

The scope of this work is necessarily limited to available data and metadata. While the topic tags provided by the UN Library are a novel contribution, the remaining data were collected from existing sources. Furthermore, our inclusion of sentiment analysis is rudimentary, providing only an average sentiment score per speech.

7 Ethical Considerations

Designed to facilitate analysis of the United Nations Security Council's debates in particular, as well as diplomatic speech and global conflict in general, we hope that our work can be a contribution to increasing transparency and insight into one of the most important decision-making bodies we have.

The UNSC-Graph contains only public data released under the CC0[9] licence. As such, there are no concerns with regard to copyright, privacy or confidentiality. Our code is Free Software, licensed under GPLv3 or later, which ensures that even derivative works will remain free.

[9] https://creativecommons.org/publicdomain/zero/1.0/

8 Acknowledgements

The first author has been partly supported by an infrastructure grant to Språkbanken Text, University of Gothenburg, for contributing to building and operating a national e-infrastructure funded jointly by the Swedish Research Council (under contract no. 2017-00626) and the participating institutions. The second author was supported by Deutsche Forschungsgemeinschaft (DFG), project (448421482) "Verläufe von Konflikten: Die Dynamik der Argumentation im UN Sicherheitsrat".

References

Fanny Badache, Sara Hellmüller, and Bilal Salaymeh. 2022. Conflict management or conflict resolution: how do major powers conceive the role of the United Nations in peacebuilding? *Contemporary Security Policy*, 43(4):547–571.

Evgeniya Bakalova and Konstanze Jüngling. 2020. Conflict over peace? The United States' and Russia's diverging conceptual approaches to peace and conflict settlement. *Europe-Asia Studies*, 72(2):155–179.

Alexander Bendix. 2022. Constructive role ambiguity and how Russia couldn't 'get away' with its 2022 Ukrainian invasion. *Central European Journal of International and Security Studies*, 16(3):108–130.

Uldis Bojārs, Roberts Darģis, Uldis Lavrinovičs, and Pēteris Paikens. 2019. LinkedSaeima: A Linked Open Dataset of Latvia's parliamentary debates. In *Semantic Systems. The Power of AI and Knowledge Graphs*, Lecture Notes in Computer Science, pages 50–56. Springer International Publishing.

Susanna Campbell and Aila M. Matanock. 2021. Weapons of the weak state: How post-conflict states shape international statebuilding. *SSRN Electronic Journal*.

Steffen Eckhard, Ronny Patz, Mirco Schönfeld, and Hilde van Meegdenburg. 2023. International bureaucrats in the UN Security Council debates: A speaker-topic network analysis. *Journal of European Public Policy*, 30(2):214–233.

Raji Ghawi and Jürgen Pfeffer. 2022. Analysis of country mentions in the debates of the un security council. In *Information Integration and Web Intelligence*, volume 13635 of *Lecture Notes in Computer Science*, page 110–115. Springer Cham.

Luis Glaser, Ronny Patz, and Manfred Stede. 2022. UNSC-NE: A named entity extension to the UN Security Council debates corpus. *Journal for Language Technology and Computational Linguistics*, 35(2):51–67.

Eero Hyvönen, Petri Leskinen, Laura Sinikallio, Matti La Mela, Jouni Tuominen, Kimmo Elo, Senka Drobac, Mikko Koho, Esko Ikkala, Minna Tamper, Rafael Leal, and Joonas Kesäniemi. 2022. Finnish Parliament on the Semantic Web: Using ParliamentSampo data service and semantic portal for studying political culture and language. In *Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop co-located with 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15, 2022*, volume 3133 of *CEUR Workshop Proceedings*, pages 69–85. CEUR-WS.org.

Juraj Medzihorsky, Milos Popovic, and Erin K. Jenne. 2017. Rhetoric of civil conflict management: United Nations Security Council debates over the Syrian civil war. *Research & Politics*, 4(2).

Stian Rødven-Eide. 2019. The Swedish PoliGraph: A semantic graph for argument mining of Swedish parliamentary data. In *Proceedings of the 6th Workshop on Argument Mining*, pages 52–57, Florence, Italy. Association for Computational Linguistics.

Cesare M Scartozzi. 2022. Climate change in the UN Security Council: An analysis of discourses and organizational trends. *International Studies Perspectives*, 23(3):290–312.

Johannes Scherzinger. 2023a. Unbowed, unbent, unbroken? Examining the validity of the responsibility to protect. *Cooperation and Conflict*, 58(1):81–101.

Johannes Scherzinger. 2023b. 'Acting under Chapter 7': Rhetorical entrapment, rhetorical hollowing, and the authorization of force in the UN Security Council, 1995–2017. *International Relations*, 37(1):3–24.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde van Meegdenburg, and Antonio Pires. 2021. The UN security council debates (version 5). Harvard Dataverse.

Mirco Schönfeld, Steffen Eckhard, Ronny Patz, and Hilde van Meegdenburg. 2019. The UN Security Council debates 1995-2017. ArXiv:1906.10969 [cs].

Minna Tamper, Rafael Leal, Laura Sinikallio, Petri Leskinen, Jouni Tuominen, and Eero Hyvönen. 2022. Extracting knowledge from parliamentary debates for studying political culture and language. In *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022), Hersonissos, Greece, May 30th, 2022*, volume 3184 of *CEUR Workshop Proceedings*, pages 70–79. CEUR-WS.org.

Astrid Van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2016. The debates of the European Parliament as Linked Open Data. *Semantic Web*, 8(2):271–281.

Maike Voss, Isabell Kump, and Paul Bochtler. 2022. Unpacking the framing of health in the United Nations Security Council. *Australian Journal of International Affairs*, 76(1):4–10.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.