Analyzing the Discourse in the UN for Crisis Response in Post-Colonial Africa

Alvan Caleb Arulandu[§]
Harvard College
Cambridge, MA
aarulandu@college.harvard.edu

Brian Zhou[§]
Thomas Jefferson High School for Science and Technology
Herndon, VA
2024bzhou@tjhsst.edu

2023 IEEE MIT Undergraduate Research Technology Conference (URTC) | 979-8-3503-0965-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/URTC60662.2023.10534975

***Abstract*—The effectiveness of international bodies such as the United Nations (UN) at addressing global crises has been debated. The updated UN General Debate Corpus (UNGDC) catalogues every speech from the UN's inception in 1946 to 2022. Using the corpus as an indicator of debate, we explore how African post-colonial states grow influence on the international stage. As these states join the UN immediately after independence, we superimpose historical events on metrics generated from UNGDC to demonstrate the corpus' robustness for forecasting shifts in international priorities. We develop a time series of relevance for each country using natural language processing methods to extract features and tokenize speeches. We conclude that the UNGD preludes intervention in multi-year violent conflicts, with insights into the efficacy of crisis resolution measures in Africa. Our results are established by computational experiments, with conditions validated by statistical significance tests.**

***Index Terms*—United Nations, UN General Assembly General Debate, UN General Debate Corpus, political communication, text as data, natural language processing**

I. Introduction

Linguistics is a flexible tool that is able to turn qualitative data into quantifiable data for data science in the field of international relations and politics. Historically, linguistics and Natural Language Processing (NLP) has been applied to a variety of applications in international relations using a variety of Corpora (compilations of text), including diplomatic documents [1], real-world conflict detection using news [2], and foreign influence from social media [3], to name a few. The growing popularity of data science and NLP approaches in social science allows the use of these Corpuses to answer important questions. One such Corpus is the UN General Debate Corpus (UNGDC) created by A. Baturo [4]. We utilize this corpus to answer questions about how post-independent states in the Global South emerge on the international stage, including their prevalence in UN discourse and distinguishing factors between African regions.

The UNGDC is comprised of speeches from the representatives of all 193 UN member states given each year in an address to the UN General Assembly (UNGA) discussing a state's stance on global issues, outlining the issues they deem most important, and laying out their national and international agendas. These speeches are important, as the UNGA is often referred to as a "barometer of international opinion" [5]. By being text-based, the UNGDC offers a promising solution to the large gap of quantitative data that existed within the social sciences just decades prior, and which continues to persist today. Notably, the broad nature of the dataset allows for a variety of quantitative studies to emerge. For example, prior studies with the UNGDC have researched how specific world leaders, countries, or even supranational bodies engage or utilize rhetoric at the UNGA, like Pakistan [6] and the speeches given by the former Prime Minister of Pakistan Nawaz Sharif [7] or the European Union [8]. Studies include a wide range of topics, including rhetoric, Islamophobia, organizational management, gender representation, language policy, and economic theory, respectively:

- Evaluating the impact of cross-cultural differences on the explicitness and persuasiveness of rhetoric between cultures at the UN [9]
- Utilizing Critical Discourse Analysis to study the portrayals of Islam and Muslims in UN speeches [10]
- Studying benign neglect in an organization's rhetoric [11]
- Highlighting language policy and ideology of UN members [12]
- Analyzing how economic scarcity between lower-income and higher-income countries influences narratives of want and wealth in speeches given at the UN [11]

Many studies have also expanded the UNGDC dataset to new parts of the UNGDC, including:

- UNSCdeb8, a corpus containing all verbatim statements of permanent and non-permanent members of the UN Security Council (UNSC) by [13]
- spaceTexts, a corpus containing all speeches related to the UN Committee on the Peaceful Uses of Outer Space including state and nonstate actors by [14]
- UNSC Afghanistan, a corpus combining quantitative and qualitative approaches to analyze dynamic topics in Afghanistan from UNSC speeches conducted by [15]

We study only the UNGDC, where new novel areas of study have been created by the expansion of the dataset. The most recent dataset created by [16] extends the original UNGDC, which covered all speeches from 1970 to 2017, to the first UNGA session, now covering from 1946 to 2022. With over

[§]Equal contribution

979-8-3503-0965-2/23/$31.00 ©2023 IEEE

10,000 speeches from more than 193 countries, the updated UNGDC is a treasure trove of linguistic data.

Uniquely, the expanded UNGDC dataset allows us to study the complete history of post-colonial states such as Nigeria or Algeria. These post-colonial states often emerged beginning in the 1950s and joined the UN the same year as their independence. All of them joined after the formation and establishment of the UN, meaning that for the entirety of the nation's existence, they have been UN members. This uniquely allows us to examine and catalog the development of these independent states as they develop economically, gain footing on the global stage, and also observe how the international community reacts and resolves turmoil in these new states.

In Materials and Methods, we discuss the composition of the UNGDC dataset, the advantages of using the UNGDC dataset for mapping the shift of a country's national priorities over time, and the specific coverage that the UNGDC dataset offers for representing lower-income countries. We use the UNGDC dataset to first develop a time series of the 'relevance' of a nation in UNGDC speeches, based on the number of references a country receives in a given UNGD. We explain the novel ways that the UNGDC dataset is able to explain trends such as UN involvement in the peacekeeping and conflict resolution process. Computational Experiments and onward detail how we extract features and tokenize speeches to conduct our experiment, as well as the conditions validated to conduct our statistical tests.

II. Materials and Methods

A. Advantages of the UNGDC Dataset

There are a few benefits to using the UNGDC. Uniquely, the UNGD provides a platform for all member states to deliver addresses, including smaller nations otherwise less represented [4], with the intention of all speeches being to justify their stance on foreign policy and persuade other UN states to adopt positions similarly [16].

This updated corpus contains over 10,000 speeches from more than 193 countries, making it the most comprehensive collection of global political discourse. This corpus can facilitate further research on a range of issues in international relations – from looking at different influences on state preferences to understanding the spread of ideas and norms in international politics.

Every September, the heads of state and other high-level country representatives gather in New York at the start of a new session of the United Nations General Assembly (UNGA) and address the Assembly in the General Debate (GD). The GD provides the governments of the almost 200 UN member states with an opportunity to present their views on international conflict and cooperation, terrorism, development, climate change and other key issues in international politics. As such, the statements made during the GD are an invaluable and largely untapped source of information on governments' policy preferences and global involvement across a wide range of issues over time. Thus, statements made during the GD can roughly approximate the relevancy of developing African nations at any particular time.

B. Feature Extraction

Before using the UNGDC, we first perform feature extraction to generate metrics for further statistical analysis. Firstly, we generate a number of characteristics describing each speech by segmenting the speech text according to stopwords provided by Python's Natural Language Toolkit [17]. Doing so, we characterize each speech according to its character count, word count, sentence count, average word length, and average sentence length.

C. Speech Tokenization

After generating these summary statistics, we tokenize each piece of speech text. In the English language, words can be modified in a number of ways while preserving meaning. For example, the words "nation", "nations", and "national" all *stem* from a common linguistic concept though they have different spellings. In order to perform an accurate linguistic analysis of UNGDC speeches, multiple variations of a single word stem must be consolidated. To execute this, we use WordNet's [18] morphological processing capabilities to isolate the stems of each word, removing affixes and other lexicographic nuances.

III. Computational Experiments

After the data has been sufficiently pre-processed in the above manner, we conduct multiple computational experiments to address the discussion of African regions in the UN.

A. Frequency Analysis

We begin by aggregating the number of mentions for each word temporally segmented by year. Sorting words by total frequency across the temporal span of the dataset, we can compute the most common words used.

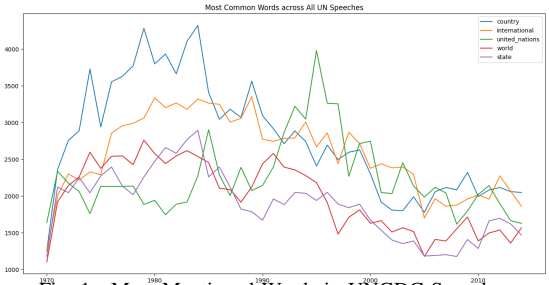

Fig. 1. Most Mentioned Words in UNGDC Speeches

The results in Figure 1 are quite intuitive, as the most popular words strongly relate to international relations and government structure. Cross-referencing these tokens with a database of countries from Python's *pycountry* package, we can compute the number of mentions over time for any given country.

From Figure 2, the most discussed countries seem to be those heavily involved in international conflicts. Using a

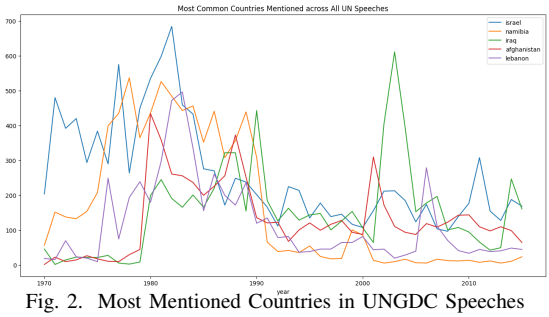

Fig. 2. Most Mentioned Countries in UNGDC Speeches

database of African regions[1] and the nations which consist them, we aggregate the net number of mentions for each region over time. From Figure 3, we see that nations in Eastern Africa seem to receive the most attention followed by Western, Northern, Southern, and Central Africa.

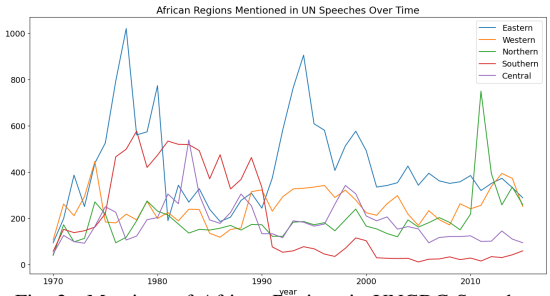

Fig. 3. Mentions of African Regions in UNGDC Speeches

B. Summary Statistics

We migrate the time series data from Figure 3 from *Python* to *R* for further analysis. We begin by transforming the time series data into a linear regression problem, viewing the quantitative response variable, the total *Count* of mentions, as a function of the quantitative explanatory variable *Year* and the categorical explanatory variable *Region*. Grouping this rearranged data by region, we compute summary statistics.

Region	min	Q1	median	Q3	max	sd
Eastern	94	311.8	360.0	523	1021	195.3
Central	46	121.0	165.5	221	539	88.1
Northern	39	149.0	172.5	211	750	104.5
Southern	11	33.2	76.5	370	577	194.8
Western	106	195.2	240.0	297	446	73.9

TABLE I
SUMMARY STATISTICS FOR AFRICAN REGION MENTIONS

From Table I, we see that the medians differ between each region, suggesting a potentially significant result.

C. Diagnosing Condition Violations

A priori, we propose and fit a multiple linear regression model with $n - 1 = 4$ indicator variables to account for the Region:

$$\text{Count} = \beta_0 + \beta_1 \cdot \text{Year} + \sum_{k=1}^{4} \beta_{k+1} I_k$$

$$\begin{aligned}\text{Count} \approx &+ 3315.017 - 1.453 \cdot \text{Year} - 237.652 \cdot I_C \\ &- 227.522 \cdot I_N - 228.022 \cdot I_S - 170.152 \cdot I_W\end{aligned}$$

From the Q-Q plot in Figure 4, the majority of the data

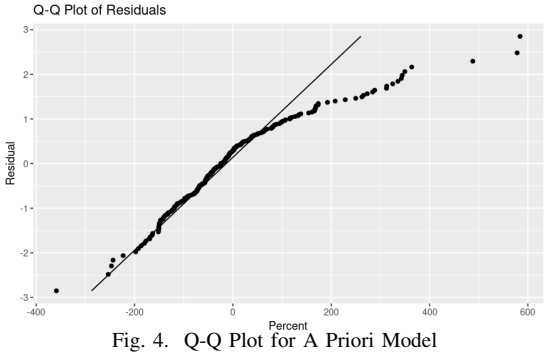

Fig. 4. Q-Q Plot for A Priori Model

adheres to normality, but there is a decent amount of concave curvature indicating a short tail on the right-hand side of the distribution. Further, we analyze the deviation in spread according to region. From Table I,

$$\frac{\max(\sigma_i)}{\min(\sigma_i)} = \frac{195.3}{73.9} = 2.64 > 2$$

This violation of normality and equal spread is enough to consider a data transformation.

D. Power Transformation

Since our data is grouped by species, we construct a diagnostic plot for computing the ideal exponent for a power transformation using the regression: $\log(\sigma) = \beta_0 + \beta_1 \cdot \log(\mu)$. Fitting this model, we find $\beta_1 = 0.538$, so we use the

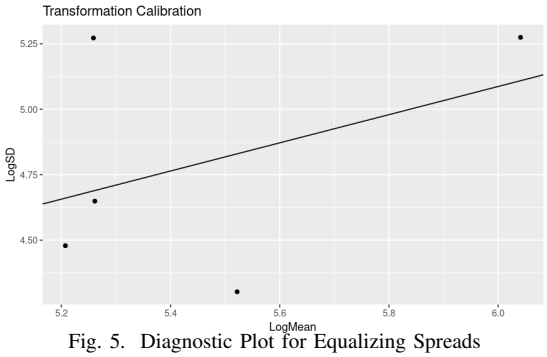

Fig. 5. Diagnostic Plot for Equalizing Spreads

following transformation:

$$x \to f(x) = x^p = x^{1-\beta_1} = x^{0.462}$$

While this transformation is similar to $\sqrt{x}$, we use this more accurate form in order to maximally adhere to normality.

[1]Regions are segmented using the UN Geoscheme for Africa from the UN M49 [19] Dependencies or overseas territories of European countries are excluded in this data, and the disputed territory of Western Sahara is included.

E. Validating Conditions

Applying $f(x)$ to both *Count* and *Year*, we fit the following regression model.

$$f(\text{Count}) = \beta_0 + \beta_1 \cdot f(\text{Year}) + \sum_{k=1}^{4} \beta_{k+1} I_k$$

$$f(\text{Count}) \approx +188.329 - 5.157 \cdot f(\text{Year}) - 5.086 \cdot I_C - 4.810 \cdot I_N - 6.082 \cdot I_S - 3.218 \cdot I_W$$

From the Q-Q plot in Figure 6, we see that while the normality

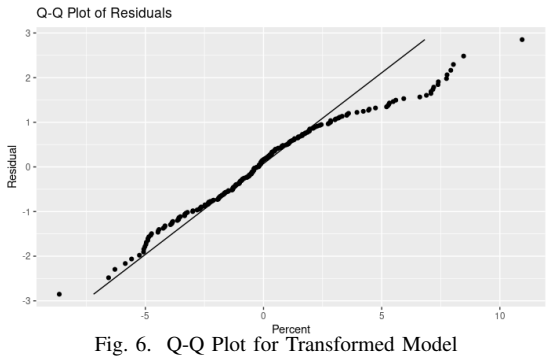

Fig. 6. Q-Q Plot for Transformed Model

condition is not perfectly satisfied, the data adheres enough to satisfy the condition. From the residual plot, we see that the

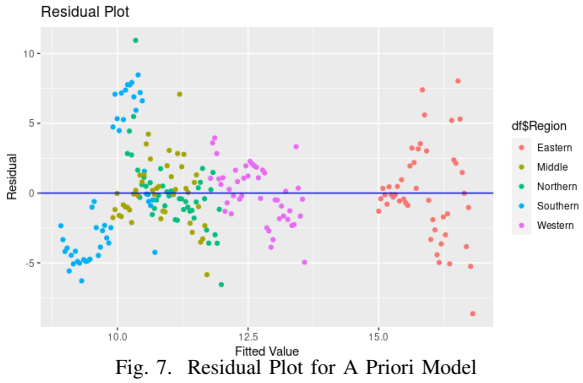

Fig. 7. Residual Plot for A Priori Model

variance across regions is sufficiently similar for the equal variance condition and note that the residuals have zero mean.

Since the entire population of UNGDC debates is used, the randomized and independent conditions are independently satisfied. We also assume that the data is approximately linear. With these conditions met, we continue to a significance test.

F. Significance Testing

We t-test the indicator weights where $H_0 : \beta_i = 0$ and $H_a : \beta_i \neq 0$ for $2 \leq i \leq 5$.

$$p_C = 1.2 \times 10^{-12} \quad p_N = 1.4 \times 10^{-11}$$
$$p_S = 2 \times 10^{-16} \quad p_W = 3.4 \times 10^{-6}$$

These resulting p-values for each indicator are all significant.

$$p_C, p_N, p_S, p_W < \alpha = 0.05$$

These indicator variable t-tests suggest that $\beta_i \neq 0$ for $2 \leq i \leq 5$, rejecting the null hypothesis and suggesting the validity of the alternate hypothesis. This is statistically significant evidence that the mean number of mentions after accounting for time is significantly pairwise different between all five African regions.

G. Complementary Correlation Analysis

Given that the conditions for multiple linear regression were not satisfied perfectly, we also analyze the Pearson correlation coefficient between each region. From Table II, most regions are weakly associated with one another. The most associated pair is Southern and Central Africa, and yet, even for this pair, $r = 0.466 < 0.5$ which is not large. This result further supports the conclusion from significance testing.

	Eastern	**Western**	**Northern**	**Southern**	**Central**
Eastern	1.000	0.228	-0.032	0.068	0.004
Western	0.228	1.000	0.293	-0.344	-0.043
Northern	-0.032	0.293	1.000	-0.176	-0.138
Southern	0.068	-0.344	-0.176	1.000	0.466
Central	0.004	-0.043	-0.138	0.466	1.000

TABLE II
PAIRWISE PEARSON CORRELATION BETWEEN REGIONS

IV. RESULTS

Understanding the timing and method by which the UN acts during crises is key to understand the effectiveness of UN solutions, such as peacekeeper deployments in Africa [20]. The UNGD is key to facilitate these understandings, as it has primarily served as a vehicle to formulate multilateral action and dialogue. Figure 8 demonstrates a spike in discourse regarding Chad during the Chad-Libyan war, and Table III demonstrates trends with UN discourse for all conflicts collected in our data. We select civil wars and conflicts that have lasted 3+ years, and compare the average mentions during the conflict with the average of the state(s) throughout its history. Two general exceptions exist. First, some states have conflicts that dominate UN discussion, such as Libya, which has a high peak due to the Libyan Civil War that overshadows other conflicts involving Libya. Second, some conflicts did not last long enough or were not severe enough to attract attention, such as Kenya and Uganda's border conflict.

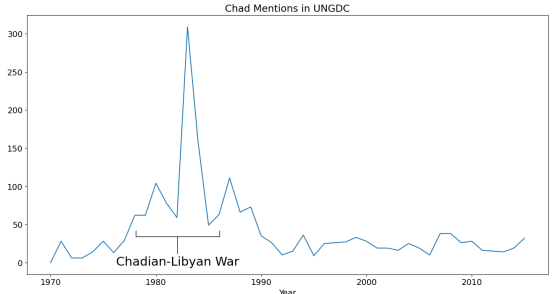

Fig. 8. Mentions of Chad during the Chad-Libyan War

State	Conflict	Period	Avg.	Time	%Δ
ETH	Ethiopian Civil War	1974−1991	28.2	**28.4**	0.71
AGO	Angolan Civil War	1975−2002	79.1	**109.5**	38.43
MOZ	Mozambican Civil War	1977−1992	41.1	**43.9**	6.81
SOM	Ethio-Somali War	1977−1978	59.4	44.5	−25.08
UGA	Uganda-Tanzania War	1978−1979	22.5	18.5	−17.78
TCD	Chadian-Libyan Conflict	1978−1987	41.8	**105.6**	152.6
UGA	Ugandan Bush War	1980−1986	22.5	15.1	−32.88
SDN	2nd Sudanese Civil War	1983−2005	50.5	43.7	−13.47
LBR	First Liberian Civil War	1989−1997	36.0	**86.2**	139.4
RWA	Rwandan Civil War	1990−1994	37.7	**112.4**	198.14
NER	Tuareg Rebellion	1990−1995	15.9	**24.0**	50.94
DJI	Djiboutian Civil War	1991−1994	11.7	10.2	-0.1282
BDI	Burundian Civil War	1993−2005	35.3	**64.2**	81.87
COD	1st Congo War	1996−1999	38.3	**107.2**	179.9
COD	2nd Congo War	1998−2003	38.3	**112.0**	192.4
LBR	2nd Liberian Civil War	1999−2003	36.0	**58.8**	63.33
LBY	Post-Civil War Violence	2011−2014	33.0	**133.0**	303.0
SSD	South Sudanese Civil War	2013−2020	50.5	**88.0**	74.26
LBY	2nd Libyan Civil War	2014−2020	33.0	**100.0**	203.0

TABLE III
MENTIONS OF A STATE DURING CONFLICT COMPARED TO ITS AVERAGE

V. CONCLUSION

Our analysis concludes that the number of mentions for Africa varies significantly between region to region based on ongoing conflicts during the time period. We identify multiple wars where discourse in the UNGD prerequisited intervention of UN actors and peacekeeping forces, especially in conflicts that lasted more than three years and with significant violence. In general, higher UN discourse in a particular conflict is correlated with the magnitude of the UN's crisis resolution response, and we aim to rigorously explore various response mechanisms in future work.

A. Future Work

Given the large amount of time series data being used, techniques such as multiple linear regression and Pearson correlation can not detect complex temporal correlations. Even if two random temporal variables are not directly correlated, there may be time-lagged cross-correlation during which changes in one variable correlate with the other after some amount of delay. Similarly, variables may temporally react in a non-Euclidean way detecting by more advanced techniques such as dynamic time warping and instantaneous phase synchrony. Measuring autocovariance could also inform the prediction of long-term mention growth trends.

Given the breadth of the UNGDC dataset, we aim to expand our experiments to consider international relations, forecast economic growth, and potentially predict global conflicts. Using statistical analysis, natural language processing, and eventually machine learning, speech analysis could prove to be an invaluable tool for foreign policy officials.

VI. ACKNOWLEDGMENTS

We would like to thank Professor Catherine Scott at the University of North Carolina Chapel Hill for her guidance.

VII. AVAILABILITY OF DATA AND MATERIAL

Our work uses Python, R, and associated packages to analyze the open-source UNGDC dataset. Please contact authors for access to the implementation.

REFERENCES

[1] M. J. Connelly, R. Hicks, R. Jervis, A. Spirling, and C. H. Suong, "Diplomatic documents data for international relations: the Freedom of Information Archive Database," *Conflict Management and Peace Science*, vol. 38, pp. 762–781, Nov. 2021. Publisher: SAGE Publications Ltd.

[2] B. O'Connor, B. M. Stewart, and N. A. Smith, "Learning to Extract International Relations from Political Context," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Sofia, Bulgaria), pp. 1094–1104, Association for Computational Linguistics, Aug. 2013.

[3] S. Kreps, "Social Media and International Relations," *Elements in International Relations*, July 2020. ISBN: 9781108920377 9781108826815 Publisher: Cambridge University Press.

[4] A. Baturo, N. Dasandi, and S. J. Mikhaylov, "Understanding state preferences with text as data: Introducing the UN General Debate corpus," *Research & Politics*, vol. 4, p. 2053168017712821, Apr. 2017. Publisher: SAGE Publications Ltd.

[5] C. B. Smith, *Politics and Process at the United Nations: The Global Dance.*, vol. 44. 2006.

[6] S. Khan, F. Ahmed, and M. Mubeen, "A Text-Mining Research Based on LDA Topic Modelling: A Corpus-Based Analysis of Pakistan's UN Assembly Speeches (1970–2018)," *International Journal of Humanities and Arts Computing*, vol. 16, pp. 214–229, Oct. 2022. Publisher: Edinburgh University Press.

[7] A. Sultan, A. Afsar, and M. A. Lashari, "Nawaz Sharif's Speeches to the United Nations General Assembly: A Corpus-Based Analysis," June 2019.

[8] M. O. Hosli and J. Kantorowicz, "The European Union in the United Nations: An Analysis of General Assembly Debates," 2022.

[9] L. Shen, "Culture and Explicitness of Persuasion: Linguistic Evidence From a 51-Year Corpus-Based Cross-Cultural Comparison of the United Nations General Debate Speeches Across 55 Countries (1970-2020)," *Cross-Cultural Research*, vol. 57, pp. 166–192, Apr. 2023. Publisher: SAGE Publications Inc.

[10] K. A. N. Al-Anbar, *The representations of Islam and Muslims at the United Nations General Assembly 2013-2016: a corpus-assisted critical discourse analysis*. phd, University of Southampton, Nov. 2018.

[11] L. McEntee-Atalianis and R. Vessey, "Using corpus linguistics to investigate agency and benign neglect in organisational language policy and planning: the United Nations as a case study," *Journal of Multilingual and Multicultural Development*, vol. 0, pp. 1–16, Feb. 2021. Publisher: Routledge _eprint: https://doi.org/10.1080/01434632.2021.1890753.

[12] L. McEntee-Atalianis and R. Vessey, "Mapping the language ideologies of organisational members: a corpus linguistic investigation of the United Nations' General Debates (1970–2016)," *Language Policy*, vol. 19, pp. 549–573, Nov. 2020.

[13] P. J. Kohlenberg, N. Godehardt, S. Aris, F. Sündermann, A. Snetkov, and J. Fall, "Introducing UNSCdeb8 (beta): A Database for Corpus-Driven Research on the United Nations Security Council," *SWP Working Paper: Research Division Asia*, p. WP 01, June 2019. Accepted: 2022-09-30T13:04:03Z Publisher: Stiftung Wissenschaft und Politik (SWP), Deutsches Institut für Internationale Politik und Sicherheit.

[14] C. Pomeroy, "spaceTexts: A new corpus of speeches in the UN committee on the peaceful uses of outer space," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, pp. 41–46, Oct. 2017.

[15] M. Schoenfeld, S. Eckhard, R. Patz, and H. van Meegdenburg, "Discursive Landscapes and Unsupervised Topic Modeling in IR: A Validation of Text-As-Data Approaches through a New Corpus of UN Security Council Speeches on Afghanistan," Oct. 2018. arXiv:1810.05572 [cs].

[16] N. Dasandi, S. Jankin, and A. Baturo, "Words to Unite Nations: The Complete UN General Debate Corpus, 1946-Present," May 2023. Publisher: Open Science Framework.

[17] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

[18] Princeton University, "About WordNet," 2010.

[19] United Nations Contributors, "Un M49: Standard Country or Area Codes for Statistical Use.," Dec 2021.

[20] A. Ruggeri, T.-I. Gizelis, and H. Dorussen, "Managing Mistrust: An Analysis of Cooperation with UN Peacekeeping in Africa," *Journal of Conflict Resolution*, vol. 57, no. 3, pp. 387–409, 2013.