# Expanding the UNSC Conflicts Corpus by Incorporating Domain Expert Annotations and LLM Experiments

**Karolina Zaczynska**
University of Potsdam
Applied Computational Linguistics Discourse Research Lab
Potsdam, Germany
zaczynska@uni-potsdam.de

## Abstract

In this work we expand the UN Security Council Conflicts corpus (UNSCon) (Zaczynska et al., 2024) on verbal disputes in diplomatic speeches in English. By including annotations of a UNSC expert, we target the problem of annotating verbal conflicts in a domain with its own culture and rules. On the one hand, we aim to catch all conflicts detected by political domain experts which as a result will be interpretable only by people with advanced political science backgrounds. On the other hand, we target linguistically marked verbalisations that are domain-independent and potentially easier to detect for language models. This balancing act resulted in a refined annotation scheme, and we re-annotate and expand the corpus size by 40% by including new debates. We perform a pilot study using a Large Language Model to include lexical markers of negative evaluation within the conflict spans, which until now were not annotated separately. Classification experiments on the conflict labels in the corpus using Transformer models demonstrate that models trained on the political domain improve the results.

## 1 Introduction

The UNSC Conflicts corpus (UNSCon) presented in our previous work (Zaczynska et al., 2024) aims to serve as a resource for understanding verbal conflicts in United Nations Security Council (UNSC) speeches. It is novel in its attempt to operationalise conflicts defined as verbal disputes and critique in a diplomatic setting, and works on disagreement detection for speeches that are mostly pre-written. We developed an annotation scheme of Conflicts including content and linguistic markers, allowing for the detection of different types of Conflicts without requiring expert knowledge of the topic. The annotations were performed by computational linguists, and had not yet been compared to those from political scientists. To address this, in this work we conduct experiments with a UN Security Council expert, identify key disagreements and suggest modifications to the annotation guidelines to improve the corpus.

Limited to debates on two topics and speeches from 2014 and 2016, UNSCon covers a restricted range of targets and periods. We expand the corpus by adding 40 new speeches on the subject *Iraq* from the years 2002, 2003, 2019, and 2020, in order to increase the diversity in topics and targets. With the expanded corpus, we perform classification experiments on Conflict types and compare them to results from the original UNSCon paper. We see that although the increasingly imbalanced label distribution between Conflicts and No Conflicts in the new dataset poses a challenge for the models, we improve scores by using RoBERTa models trained on argumentation and the political domain.

Detecting lexical markers of negative evaluation within Conflict spans is a crucial part of annotating these spans and is required for certain Conflict labels. Currently, annotations are applied to Elementary Discourse Units (EDUs), which are typically sentences or clauses. These annotations define Conflict types within the EDUs but do not specify the lexical markers themselves. To enhance the corpus' granularity, we conduct a pilot study using Large Language Models (LLMs) to identify the lexical markers inside the Conflict spans (EDUs) and categorise different types of lexical markers that indicate negative evaluation.

To summarise our contributions, we expand the corpus on two levels, qualitatively and quantitatively:

- We aim to improve the quality of annotations and the annotation scheme by incorporating suggestions made by an UNSC domain expert (§3).
- We expand the corpus: (1) by incorporating speeches from an additional topic (§4), and



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 343–357
July 5–10, 2020. 

(2) by incorporating automatically detected lexical markers of negative evaluation within the Conflict text spans using an LLM (§5.1 and 6.1).

- We provide new classification experiments for Conflict type detection on the refined and expanded UNSCon, compare the results with those obtained from the original corpus, and demonstrate improvements testing on RoBERTa models trained on similar tasks and domains (§5.2 and 6.2).

The updated dataset and the code for experiments are available in our GitHub repository.[1]

The remainder of the paper is structured as follows: First, we present related work and detail the annotation scheme for Conflict types as defined in Zaczynska et al. (2024) (§2). Next, we describe the annotation experiments conducted with a political scientist (§3) and the updated Conflicts annotation scheme based on identified disagreements. Then, we introduce our expanded dataset with new annotation guidelines and the additional speeches included (§4). We outline the experiments and classification setups (§5), discuss the results (§6), and, finally, draw conclusions (§7).

## 2 Background

In our former work presenting the UNSCon (Zaczynska et al., 2024), we define Conflicts as verbal disagreements or critique directed at someone present at the UNSC debate, without necessarily referring to a military or physical conflict. There are different types of Conflict:

**(1) Negative Evaluations (NegE)** describe Conflicts where the speaker directly criticises another country (DIRECT NEGE). Speakers can also criticise an intermediate entity serving as a proxy instead of directly targeting another country (INDIRECT NEGE). Below is an example from a speech given on Ukraine after a resolution criticising a referendum planned in Crimea was vetoed by the Russian Federation. It starts with a direct critique on Russia's voting behaviour (labelled with the Conflict type DIRECT NEGE) and continues with a critique of the referendum that Russia supports (labelled as INDIRECT NEGE):

(1) Russia's decision to veto the resolution is therefore profoundly unsettling. – DIRECT NEGE
The referendum to be held tomorrow in Crimea is dangerous and destabilizing. – INDIRECT NEGE
It is unauthorized and invalid. – INDIRECT NEGE
(S/PV.7138, Australia)[2]

**(2) Challenge and Corrections (CC)** describe Conflicts where a speaker accuses another one of lying (CHALLENGE) and where a speaker provides a correction to that allegedly false statement (CORRECTIONS). The next example is taken from a speech in which the speaker from the Russian Federation is addressing accusations made by the United States:

(2) The Permanent Representative of the United States blamed Russia for illegally pursuing its ambitions. – CHALLENGE
That does not apply to us; – CORRECTION
it is a phrase taken from the foreign policy arsenal of the United States.
(S/PV.7138, Russian Federation)

For an EDU to be a Conflict, it must be possible to identify a target (addressee) of the critique by examining the speech. The annotation scheme specifies a set of target types for the Conflict, along with the specific countries being targeted. The UNSCon includes 87 speeches from debates discussing two topics: the *Ukraine* conflict, and the *Women, Peace and Security agenda* (WPS) focusing on gender (in)equality and crimes committed during peace keeping missions. The annotation spans are Elementary Discourse Units (EDUs) based on Rhetorical Structure Theory (Mann and Thompson, 1988). EDUs are usually sentences or clauses.

The work on the UNSCon is based on transcriptions of meetings in the UNSC (Schoenfeld et al., 2019), which serve as a foundation for various analyses in linguistics, computational linguistics, and political science. For example, Anisimova and Zikánová (2024) examine how diplomats convey evaluative speech using appraisal theory (Martin and White, 2005) for their analysis. Other studies focus on extracting country mentions in UNSC discussions using Wikidata for Named Entity Linking

[1] https://github.com/linatal/Expanding_UNSCon

[2] All examples are taken from the UNSCon and labelled with the original debate-id and country name the speaker represents.

(Glaser et al., 2022) and Named Entity Recognition (Ghawi and Pfeffer, 2022). Network analyses have also been conducted on UNSC topics from Afghanistan debates (Eckhard et al., 2021). Scartozzi (2022) look at discourse related to climate change in the UNSC.

Reinig et al. (2024) created a new resource of German parliamentary debates, annotated with fine-grained speech act types distinguishing between cooperation and conflict communication. Focusing on discourse in political debates around the US election 2016, Visser et al. (2020) annotated argument relations using the relation classes Inference, Conflict, and Rephrase. Focussing on dialogues they use the term Conflict differently than in the UNSCon, indicating incompatible propositions.

## 3 Evolution of the Annotation Scheme based on Domain Expert Annotations

In this section, we compare parallel Conflict annotations of the UNSCon speeches made by a UN Security Council expert with the original ones made by computational linguists. The analysis is the basis for the refined annotation scheme we present in the following sections. We first present the Inter-Annotator Agreement (IAA), along with some general observations, followed by a detailed analysis of the most common disagreements in the annotations.

### 3.1 General Observations and IAA

For the annotation experiments, we provided the political domain expert with annotation guidelines and used the pre-segmented raw texts from the original dataset.[3] Annotations were performed on all 87 speeches. Since we are working with potentially overlapping span annotations, we calculated IAA between the UNSCon annotations in the original corpus and the domain expert's annotations using unitising Krippendorff's alpha (Krippendorff, 2004). For INDIRECT versus DIRECT NEGE Conflict types versus NO CONFLICT, the IAA is 0.3, and for Targets, it ranges from 0.32 to 0.37. For CHALLENGE versus CORRECTION versus NO CONFLICT, the IAA is 0.37. The agreement is lower than what Zaczynska et al. (2024) reported for their experiments but still moderate, considering that their annotators received training during weekly meetings to resolve borderline cases. In contrast, our annotator conducted annotations mainly based on the provided guidelines without additional training.

In the original dataset, Conflicts usually span entire sentences, with a few exceptions. We observe that the political scientist annotator often chose to annotate individual propositions rather than full sentences as Conflict spans. When both NEGE and CC were applicable, the original UNSCon annotations preferred CC (which is according to the annotation guidelines), while the political domain expert frequently chose NEGE instead of CORRECTION. Generally, the political domain expert often labelled CORRECTION differently: Of the 148 EDUs labelled as CORRECTION in the original dataset, 17% (35 EDUs) were classified as NegEval by the political domain expert, and 21% (31 EDUs) were even marked as NO CONFLICT. Beyond that, there are similar disagreements to those identified by Zaczynska et al. (2024), such as interchanging INDIRECT with DIRECT NEGE. Of the 424 EDUs labelled as INDIRECT NEGE in the original dataset, 13% (56 EDUs) were classified as DIRECT NEGE by the political scientist. The following subsections address the disagreements we found between the annotations.

### 3.2 Diplomatic Phrasing

The choice of words is important in diplomacy; a restrained vocabulary allows nuanced control when agreeing or disagreeing with others to prevent unintended enthusiasm or offence (Stanko, 2001).[4] Thus, it is not surprising the political domain expert annotated Conflicts based on diplomatic rules, which the UNSCon did not include. For example, the sentence in bold below was marked by the domain expert as DIRECT NEGE due to its suggestion of a complaint about the Council's delayed discussion.[5] In contrast, productive meetings would be indicated by phrases like "it is a good opportunity [...]".

(3) The United States deeply appreciates the support from our colleagues around the table and from the many States that have called for a peaceful end to the crisis in Ukraine. This is, however, a sad and remarkable moment. **It is the seventh time that the Security Council**

[3] Both available online: https://github.com/linatal/UNSCon

[4] Some studies suggest this ambiguity is used strategically to achieve objectives (Bach et al., 2025; Scott, 2001).

[5] **Emphases** here and in the following examples are by paper's author.

**has convened to discuss the urgent crisis in Ukraine.** The Council is meeting on Ukraine because it is the job of this body to stand up for peace and to defend those in danger. (S/PV.7138, United States)

To maintain a clear linguistic operationalisation of Conflicts in the corpus, we chose not to include these implicit Conflicts. Consequently, this example shows, that the UNSCon may not contain all sentences marked with this type of critique, also in the updated version.

### 3.3 Instructions

A similar subtle critique as in (3) is present in the next example as an instructive formulation. Here, the representative of China communicates that more time should have been given before voting on the solution. This was not annotated in the original UNSCon, but it was marked by the political domain expert as DIRECT NEGE:

(4) We believe that the Security Council **should have had ample time** for further consultation to maximize our efforts to seek agreement and forge consensus to the largest extent possible. (S/PV.7643_spch008, China)

This example highlights the challenge of distinguishing between critical directives and, conversely, motivating or positively suggesting something in political speech.

Examining the domain expert annotations, we found differing assessments of whether instructive words carried conflict-related meaning. The next example includes "must", which caused the domain expert to annotate the sentence as Conflict, given its formulation as a strong demand implying criticism of Russia. The repetition reinforces this effect.

(5) Russia **must** pull back its forces to their bases and decrease their numbers to agreed levels. It **must** allow international observers access to Crimea. It **must** demonstrate its respect for the sovereignty and territorial integrity of Ukraine, [...]. It **must** engage in direct dialogue with Ukraine, as Ukraine has repeatedly requested, [...]. (S/PV.7138_spch012, Australia)

In a study by Gruenberg (2009) on the language used in UNSC resolutions, a small taxonomy of instructive words is presented, ranking them from weakest to strongest (see Figure 1). For instructive sentences, we use the hierarchy provided by Gruenberg (2009) to update the Conflict annotations accordingly, since it resembles the assessments of our domain expert. Annotators are now advised to consider marking instructive words stronger than "recommend" as NEGE, noting that this should be assessed case-by-case. In the range of instructive words shown in Fig. 1 we can rank "must" between "request" and "urge".

| Emotive Words From Weakest to Strongest |
|---|
| Concerned |
| Grieved |
| Deplored |
| Condemned |
| Alarmed |
| Shocked |
| Indignant |
| Censured |

| Instructive Words From Weakest to Strongest |
|---|
| Decide |
| Call upon |
| Recommend |
| Request |
| Urge |
| Warn |
| Demand |

Figure 1: Range of emotive and instructive words from weakest to strongest taken from Gruenberg (2009).

### 3.4 Emotive Words

The Security Council employs a diverse vocabulary to express its institutional stance on different entities. While in the UNSCon the next two sentences were not annotated as Conflict, the domain expert chose DIRECT NEGE and explained this with the UK representative's decision to use "condemn". At the same time, we saw that sentences including "call upon" or 'urge" were not annotated. Gruenberg (2009) categorised emotive words by intensity (see Figure 1), where "condemned" falls in the middle range.

(6) The United Kingdom **condemns** the abduction at gunpoint and public parading of an OSCE Vienna Document inspection team and its Ukrainian escorts. (S/PV.7138, United States)

Similar to instructive words, for the improved UNSCon annotations, we include the hierarchy of emotive words by Gruenberg (2009) into the annotation guidelines and recommend considering the annotation of Conflicts based on emotive words that are similar or stronger than "condemned".

### 3.5 Sarcasm and Rhetorical Questions

From what we observed in the corpus, rhetorical questions and sarcasm often indicate a confrontational tone of statements in the UNSC speeches (and were accordingly annotated as Conflict by

the UNSC expert), but were not annotated in the original corpus because they did not fit into existing Conflict type annotation rules. Another reason for including these types of utterances in the Conflict annotation scheme is informed by literature from political science, which discusses how sarcasm and humour are used in diplomacy to provoke, undermine discourse, or argue (Brassett et al., 2021; Chernobrov, 2023). The next example shows no lexical marker of negative evaluation, but the Russian representative uses a sarcastic tone to criticise other Council speakers. The political domain expert annotator labelled both annotations as DIRECT NEGE.

(7) **Some colleagues** today have achieved **high levels of rhetoric**. I must mention that the Ukrainian colleague nevertheless went far beyond anything permissible. [...]. (S/PV.7138_spch020, Russia)

In the example, the use of "some colleagues" can be interpreted as a defamatory reference to someone in the room; using "high levels of rhetoric" is a confrontational way of criticising others' speeches. It is sarcastic since the literal meaning is positive, but pragmatically it is intended to express a critique. In the next example, the representative of Lithuania uses a rhetorical question to criticise the statements given by the Russian representative, framing separatist groups as "peaceful protesters". Again, this sentence was marked by the domain expert, but not in the original dataset.

(8) A few days ago, a Ukrainian helicopter was downed by a rocket-propelled grenade, hardly a weapon so-called peaceful protesters - as labelled by the Russian side - can buy at the local corner market. **That certainly does not sound like the implementation of Geneva agreement by the separatists and their state sponsors?** (S/PV.7165_spch016, Lithuania)

Since we encountered several such instances, we added a new label FIGURATIVE LANGUAGE (FIGL) to the Conflict guidelines, covering sarcasm (saying something opposite of what is meant) and rhetorical questions (asking a question not to receive an answer, but to make a point or convey irony). The Appendix in section A provides more detailed guidelines for detecting sarcasm and rhetorical questions.

### 3.6 Cultural Differences in expressing Conflict

Conflicts from certain countries are more subtle compared to others, often avoiding direct naming of the addressee of the critique. Requiring lexical markers and identifying a target may result in missing Conflicts in less confrontational speeches. Some statements were marked as NEGE by the UNSC expert when the targeted country in the Council was inferred through background knowledge of the discourse. However, when they cannot be determined by the speech alone, they are not in the original corpus.

In the next example, the last sentence is a candidate for Conflict and was marked by the political scientist, but the speech is so implicit in not naming a target that it is unclear whether it refers to a country or a non-governmental group, making it difficult to determine the conflict type. Therefore we decided not to include this and similar Conflicts in the dataset, even if it means losing some conflict statements.

(9) We are troubled in particular by the continuing violence and aggressive provocations by illegal armed groups, including the seizure of key public buildings and the recent assassination attempt against the Mayor of the eastern city of Kharkiv. **All provocative actions and hostile rhetoric aimed at destabilizing Ukraine must cease immediately.** (S/PV.7165_spch010, Korea)

We also observed that some countries use more sarcasm and rhetorical questions than others. These cultural differences in communication were not included in the previous annotation scheme, which we now have addressed by including these as Conflict types.

## 4 Corpus Extension by Size

In this section we describe the extension of the UNSCon not only through applying the refined annotation guidelines to existing speeches but also by including new speeches from new debates.

To broaden the scope of the UNSCon, which concentrates on Ukraine and the WPS agenda, we included debates on Iraq. These debates focus on an (imminent) military conflict in Iraq, highlighting a crisis in international relations and the formation of opposing factions within UNSC countries — one supporting the military operation (including the US and Great Britain), and another opposing it

| | #EDUs | |
|---|---|---|
| **Conflict Type** | **UNSCon** | **extended** |
| Direct NegE | 771 | 1621 |
| Indirect NegE | 501 | 516 |
| Challenge | 101 | 138 |
| Correction | 128 | 214 |
| Sarcasm | - | 52 |
| Rhetorical Question | - | 120 |
| Conflict | 1501 | 2642 |
| No Conflict | 4497 | 7162 |
| **Sum** | **5998** | **9804** |

Table 1: UNSCon statistics original and updated version.

(Russian Federation, France, and others). We also included 2019 and 2020 debates on Iraq covering topics like the formation of a new Iraqi government, the violent response of the previous Iraqi government to demonstrations, and the threat posed by Islamic State (IS) terrorist groups in Iraq. Having a broader range of topics not directly related to military conflicts is more representative of other UNSC discussions, though they have a smaller total amount of Conflicts.

### 4.1 Corpus Statistics Expanded UNSCon

The corpus extension was carried out by the paper's author. For the EDU segmentation of the newly added speeches, we used Kamaladdini Ezzabady et al. (2021)'s MELODI system, which is available as part of the GitLab project page for their DisCut22 Discourse Annotator Tool.[6] We chose this system due to its accessibility and because it reported an f1-score of over 0.9 on the EDU segmentation task within the DISRPT2021 shared task. We expanded the corpus by segmenting and annotating it further, increasing the number of Elementary Discourse Units (EDUs) by 39%, and the number of Conflict annotations by 43%, resulting in a total of 9,806 EDUs (before: 5,998), and 131 speeches from 14 different debates (previously 87 speeches from 6 debates). The updated corpus now includes Conflicts originating from speeches delivered by 23 different countries (before: 21) and these speeches are targeted at 13 different countries (before: 5). Table 1 shows a more detailed comparison of the label distribution between the two versions of UNSCon.

[6] https://gitlab.irit.fr/melodi/andiamo/discoursesegmentation/discut22

We observe a greater imbalance between Conflicts and No Conflicts, with a tendency towards more No Conflict EDUs compared to the original version. With the inclusion of debates on additional topics, such as the spread of IS, we see that most countries criticise IS rather than each other, which is why they were not annotated as Conflicts. This may pose a challenge for classifiers; however, we view this as a more accurate representation of the general nature of speeches given at the UNSC, as the previous dataset predominantly consisted of highly controversial debates, mostly centred on the Ukraine crisis.

### 4.2 Inter-Annotator Agreement Expanded UNSCon

To evaluate the extension of the corpus done by the paper's author and the refined annotation guidelines, we had a second annotator (a computational linguistics student) annotate over 10% of the extended corpus. We selected speeches mainly from the new topic Iraq, as well as those containing instructive and figurative language. For NEGE, Cohen's Kappa is 0.71, which is slightly less than Zaczynska et al. (2024) report. For Krippendorff's Alpha (unitising) we report 0.6 for NEGE (two labels), 0.57 for Target Council (six labels), 0.59 Target Intermediate (six labels), and 0.65 for Country Name (nine labels). For Challenge Type (two labels), we report an Krippendorff's Alpha of 0.68, Target Challenge (five labels) 0.64, Country Name (eight labels) 0.64. For NEGE and CC, it appears that when there is agreement on the position and conflict type, agreement regarding the targets is similar to the previous labels. However, for FIGL, we observe a different pattern. For FIGL Type, we see a reasonable agreement with 0.61, but a lower agreement for the Targets (0.27 for Target Type and 0.25 for Country Type). This indicates a challenge in including this new Conflict type, as neither Sarcasm nor Rhetorical Questions necessarily clearly verbalise a target of the critique. However, with only a few instances of annotation for FIGL (166 EDUs), these observations should be taken cautiously.

## 5 Experiments

The next section outlines our setups for two sets of experiments: first, a pilot study on half of the dataset to incorporate lexical marker annotations for UNSCon, and second, an experiment utilising

Transformer models for fine-tuning on the Conflict type classification task.

### 5.1 Expansion of Conflicts with Lexical Markers

We perform a pilot study on using LLMs to extract the spans that include lexical markers of negative evaluation. Additionally, we let the LLM categorise the extracted lexical marker according to categories that are expanded and are more structured compared to the original guidelines.

- "Adjectival_Attribution": Adjectival attributions like *bad, dreadful, worrying*)
- "Noun": Nouns with a negative connotation (e.g., *traitor, annexation*)
- "Adverb": Adverbs that intensify criticism (e.g., *poorly, even, only*)
- "Verb": Verbs with a negative connotation (e.g., *infiltrating, invading*)
- "Negation_Phrase_or_Quantifier": Negation phrases and quantifiers (e.g., *not at all, not a single*)
- "Evaluative_Pattern": Recognisable evaluative patterns (e.g., *It is unfortunate that...*, *There is something worrying about...*)
- "Instructive_Words": Strong instructive words (e.g., *urge, must, warn, demand*)
- "Emotive_Words": Strong emotive words (e.g., *condemned, armed, shocked*)

For our pilot study, we use GPT4o (OpenAI, 2024) to annotate about half of the dataset (5,049 EDUs). Other open source models (llama-3.3-70b-versatile[7], gemma2-9b-it[8]) we tested did not produce satisfactory output. This might be due to the relatively complex task which consists of three steps: first, detecting if there are one or more lexical markers, second, categorising them, and third, extracting the substring(s) from an EDU. The final prompt we used for the experiment is provided in the Appendix B.

### 5.2 Classification Setup

We classify conflicts from diplomatic sources according to four distinct subtasks:

- 2-class setup, no FIGL: For comparability with the former classification setup, which did not include figurative language. We exclude the FIGL label for this setup.
- 3-class setup, no FIGL: For comparability with former classification setup, models should label each EDU choosing from one of the three categories: No Conflict, NEGE, CC.
- 4-class setup: models should label each EDU choosing from one of the four categories: No Conflict, NEGE, CC, FIGL.

We did not include more fine-grained classification on Conflict labels because of the performance drop we see for the 3 and 4-class setup (see section 6).

We test the following models on the UNSCon-extended for the classification tasks: We evaluated the best performing system reported in Zaczynska et al. (2024), namely RoBERTa-argument[9], which was trained on a variety of text types for binary classification tasks of arguments versus non-arguments. Given that none of the formerly tested models were trained on the political text domain, we additionally evaluated the following two models: PolicyBERTa-7d[10] (henceforth: RoBERTa-policy) is trained for topic detection based on the Manifesto Project, a project that collected election manifestos to study parties' policy preferences. Additionally, we also tested ArgumentMining-EN-ARI-AIF-RoBERTa_L (Ruiz-Dolz et al., 2021)[11] (henceforth: RoBERTa-relations) a model trained on a dataset tailored to a more fine-grained task than binary argumentation detection, specifically focusing on Argument Relation Mining, which involves classifying text into Inference, Conflict, and Rephrase relations. This model was trained on the datasets US2016 (Visser et al., 2020), containing annotated television debates and social media reactions to the US campaign in 2016, and on QT30 (Hautli-Janisz et al., 2022), a corpus focused on arguments and conflicts in Broadcast Debate. We follow the previous configurations as detailed in Zaczynska et al. (2024)(learning rate 1e-5, batch size of 32, with 2 training epochs and a weight decay of 0.01). We train the classifier to assign labels

[7] https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

[8] https://huggingface.co/google/gemma-2-9b-it

[9] https://huggingface.co/chkla/roberta-argument

[10] https://huggingface.co/niksmer/PolicyBERTa-7d

[11] https://huggingface.co/raruidol/ArgumentMining-EN-ARI-AIF-RoBERTa_L

for EDUs. All scores reported for the models are the result of 10-fold cross-validation.

## 6 Results and Discussion

### 6.1 Linguistics Markers

We perform a comparative analysis of the categories and lexical markers identified in a test set of 134 EDUs, using output from GPT4o and comparing it with another LLM, Gemini 2.0 Flash (Gemini). For calculating Cohen's Kappa, we ignore the text span length and focus solely on comparing the lists of categories assigned to each EDU by the two systems. For categories, we observe an average Cohen's Kappa of 0.45. In our multi-label setting, where multiple lexical marker annotations can exist per EDU, Cohen's Kappa is only partially appropriate because it allows the comparison of only one single point with another. We therefore also provide set comparison using the Jaccard index, where for each EDU, we compare all lexical markers and categories found for one EDU from Gemini against GPT4o as sets of strings and extract an overlap measure. For lexical marker categories, we observe an average Jaccard index of 0.63, and for extracted strings 0.59. Comparing the two outputs qualitatively, we see similar results regarding what is identified as a lexical marker of negative evaluation in the text; however, the chosen span of annotation differs. While GPT4o extracts phrases (for example, *camp of war in opposition to the United Nations and its Charter*), Gemini extracts individual words (*war, aggression, opposition*), and therefore, this also affects the categorisation: Because GPT4o focuses on phrases, it more frequently selects "Recognisable evaluative pattern" (*do its bidding -> Recognisable evaluative pattern*, *Negative verb*), whereas Gemini selects more specific word types (*make, do, bidding -> Verbs with a negative connotation, Strong instructive words*). Thus, while there is significant overlap of the chosen regions within the EDUs as being identified as lexical markers between both model outputs, the different spans negatively impact the IAA.

Looking at the distribution of lexical marker categories found in the annotated dataset we see that for all Conflict types the most prominent lexical markers are nouns with a negative attribution, followed by verbs (see Figure 2). A list of most frequent words (lemmatised using SpaCy library (Honnibal et al., 2020)) is in the Appendix C.

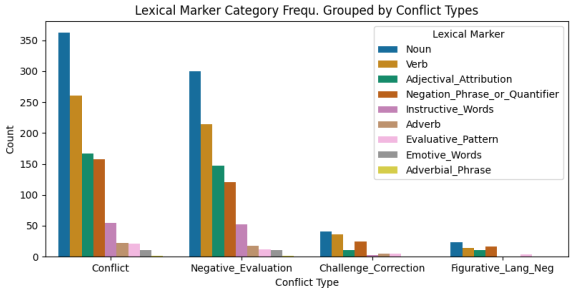


Figure 2: Frequency of found Lexical Marker Categories per Conflict Types.

### 6.2 Model Performance Classification

In Table 2, we present the classification results for the 3-class and 4-class setups. In our classification experiments on Conflict types using various RoBERTa-based models, we observe that for the binary setup (excluding FIGL, as it is absent from the old dataset), the results reported in Zaczynska et al. (2024) outperform our models fine-tuned on the new dataset. They report an f1-macro score of 0.74, whereas we achieve a best result of 0.70 for RoBERTa-relations. Comparing the performance of RoBERTa-argument on the old dataset with the new one, we note slightly better results for the binary and 3-class setups in the former (f1-macro 0.48 versus 0.45). We hypothesise that, although it offers more training instances, this is due to the increased label imbalance in the new corpus.

Comparing the results on our new dataset, RoBERTa-policy performs slightly better than RoBERTa-argument, although still lower than RoBERTa-relations. RoBERTa-policy was trained on topic detection using party manifestos, which are more similar to diplomatic texts than the diverse texts RoBERTa-argument was trained on.

Examining the 3-class setup (labels NegE, CC, or No Conflict), RoBERTa-relations again yields the best scores, outperforming RoBERTa-argument fine-tuned on the old dataset. We think that the good performance of RoBERTa-relations is due to the fact that it was trained on fine-grained Argument Relations classification and on political debates. The classification results thus suggest that domain-specific training — even when not on diplomatic texts but more broadly on political domains — enhance performance on Conflict classification tasks.

| | UNSCon extended | | | orig. UNSCon |
|---|---|---|---|---|
| | RoBERTa$^{\text{argument}}$ | RoBERTa$^{\text{policy}}_{\text{topics}}$ | RoBERTa$^{\text{argument}}_{\text{relations}}$ | RoBERTa$^{\text{argument}}$ |
| 2-class setup (Conflict / No Conflict, without FigL) | | | | |
| precision | 0.72 | 0.72 | 0.73 | **0.78** |
| recall | 0.68 | 0.68 | 0.69 | **0.78** |
| f1-macro | 0.70 | 0.69 | 0.70 | **0.74** |
| accuracy | 0.78 | 0.79 | **0.79** | 0.78 |
| 3-class setup (NegE / CC / No Conflict) | | | | |
| precision (macro avg) | 0.45 | 0.45 | 0.64 | **0.72** |
| recall (macro avg) | 0.45 | 0.45 | 0.48 | **0.76** |
| f1-macro | 0.45 | 0.45 | **0.51** | 0.48 |
| accuracy | 0.77 | 0.78 | **0.78** | 0.76 |
| 4-class setup (FigL / NegE / CC / No Conflict) | | | | |
| precision (macro avg) | 0.34 | 0.58 | **0.62** | N/A |
| recall (macro avg) | 0.34 | 0.33 | **0.42** | N/A |
| f1-macro | 0.33 | 0.34 | **0.47** | N/A |
| accuracy | 0.77 | 0.76 | **0.77** | N/A |

Table 2: Classification results of the (1) 2-class setup: comparing the reported performance of the best model from Zaczynska et al. (2024) on the original UNSCon, and different RoBERTa-based models fine-tuned on the extended corpus, excluding FigL for comparability; (2) 3-class setup: comparing results reported on the original UNSCon fine-tuned on RoBERTa-argument with fine-tuned models on the new corpus, again excluding FIGL label; and (3) 4-class setup: comparing fine-tuned models on the new corpus including FIGL label.

## 7 Conclusion

This paper presents an extended version of the UNSC Conflicts Corpus as introduced by Zaczynska et al. (2024), by expanding both the annotation guidelines and corpus size, and incorporating more detailed annotations of lexical markers of Conflicts using an LLM. Working with diplomatic texts, and being annotated by computational linguists, we provide a detailed evaluation of political scientist annotations on the corpus and discuss identified disagreements. Annotating communicative phenomena in language within NLP, especially in a domain with its own culture and rules such as the diplomatic setting, presents a balancing act regarding annotation guidelines. One must choose between creating guidelines that target diplomatic language usage only interpretable by people with advanced political science backgrounds, and linguistically marked verbalisations that are relatively domain-independent and possible to pick up on by NLP classifiers. We refined the annotation scheme and kept both the original notion of a mandatory lexical verbalisation of Conflict, and also included Conflict labels that might need cultural knowledge to detect, like figurative language.

Our classification experiments on Conflict types using Transformer models show that integrating a model trained on a similar task and domain improves the performance. Despite this, the results indicate that smaller Conflict types like CHALLENGE CORRECTION (CC) (which involves detecting when someone claims another speaker is lying, and the correction of this alleged lie), and FIGURATIVE LANGUAGE (FIGL) (which includes sarcasm and rhetorical questions) require more data to achieve satisfactory outcomes. Looking at the classification results for each Conflict label, we observe that all models struggled to accurately classify less frequent classes. In addition to the small number of training samples, this also may be attributed to the inherent difficulty of the task. Detecting FIGURATIVE LANGUAGE, for instance, remains a challenge in NLP (Liu et al., 2022). However, training on dedicated task-specific datasets might enhance performance (Sanchez-Bayona and Agerri, 2024). For future work we will conduct a further qualitative analysis of the lexical markers and types extracted by the LLM and will expand the experiments to the full dataset. Additionally, we plan to broaden the current limited list of emotive and instructive words by Gruenberg (2009) into a larger taxonomy, using the list of lexical markers found in the experiments by the LLM, including terms expressing negative assessments found in the speeches.

## Limitations

The study relies on annotations from a single political scientist, and gold annotations for the new UNSCon dataset was also done by one annotator, which may introduce bias into the analysis of annotation disagreements. Regarding our observations on cultural differences in expressing Conflicts, we must note that some speeches are originally given in other languages and then translated into English by UN personnel. Although the UNSC employs institutional mechanisms to ensure high-quality translations (such as monitoring programs, terminology, and proofreading),[12] these translations might introduce some bias or alter meanings or tone, potentially affecting the annotation of Conflicts. This issue may be particularly relevant for fine-grained annotations of sarcasm. Replicating the study in a language other than English might yield different Conflict annotations.

## Acknowledgments

The paper was supported by the Deutsche Forschungsgemeinschaft (DFG), project "Trajectories of Conflict: The Dynamics of Argumentation in the UN Security Council“ (448421482) and the German Academic Exchange Service (Deutscher Akademischer Austauschdienst) by a research grant for doctoral students. We thank our annotators Dr. Antonio Pires and Dietmar Benndorf for their work and their valuable feedback on the annotation guidelines. We thank Costanza Rasi for helping with the automatic EDU segmentation. We thank Prof. Manfred Stede and Dr. Peter Bourgonje on their helpful feedback on the paper.

## References

Mariia Anisimova and Šárka Zikánová. 2024. Attitudes in diplomatic speeches: Introducing the CoDipA UNSC 1.0. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 17–26, Torino, Italia. ELRA and ICCL.

Parker Bach, Carolyn E Schmitt, and Shannon C McGregor. 2025. Let me be perfectly unclear: strategic ambiguity in political communication. *Communication Theory*, page qtaf001.

James Brassett, Browning , Christopher, , and Muireann O'Dwyer. 2021. EU've got to be kidding: Anxiety, humour and ontological security. 35(1):8–26.

Dmitry Chernobrov. 2023. Strategic humor and post-truth public diplomacy. Discussion paper, ARRAY(0x56430ed8ae38). © 2023.

Bernard Comrie and Jerrold Sadock. 1974. Toward a linguistic theory of speech acts. *Philosophical Quarterly*, 26(104):285.

Martina Ducret, Lauren Kruse, Carlos Martinez, Anna Feldman, and Jing Peng. 2020. You don't say... linguistic features in sarcasm detection. In Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021*, Collana dell'Associazione Italiana di Linguistica Computazionale, pages 171–177. Accademia University Press. Code: Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021.

Steffen Eckhard, Ronny Patz, Mirco Schönfeld, and Hilde van Meegdenburg. 2021. 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*, pages 110–115, Cham. Springer Nature Switzerland.

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.

Justin Gruenberg. 2009. An analysis of united nations security council resolutions: Are all countries treated equally? *Case Western Reserve Journal of International Law*, 41(2):513.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. 2021. Multi-lingual discourse segmentation and connective identification: MELODI at disrpt2021. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[12] https://www.rferl.org/a/UN_Interpreters_Make_Sure_Nothing_Is_Lost_In_Translation/1995801.html

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. 38(6):787–800.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

J. R. Martin and P. R. R. White. 2005. *The Language of Evaluation*. Palgrave Macmillan UK.

Antonio Jesús Moreno-Ortiz and María García-Gámez. 2022. Corpus annotation and analysis of sarcasm in twitter: #CatsMovie vs. #TheRiseOfSkywalker. *Atlantis. Journal of the Spanish Association for Anglo-American Studies*, pages 186–207.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Hannah Rohde. 2006. Rhetorical questions as redundant interrogatives. *UC San Diego: San Diego Linguistic Papers*.

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barbera, and Ana Garcia-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation. *ArXiv*, abs/2404.07053.

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.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde van Meegdenburg, and Antonio Pires. 2019. The un security council debates 1992-2023. *Preprint*, arXiv:1906.10969.

Norman Scott. 2001. Ambiguity versus precision: The changing role of terminology in conference diplomacy - diplo resource. In *Langauge and Diplomacy*.

Stephen Skalicky and Scott Crossley. 2018. Linguistic features of sarcasm and metaphor production quality. In *Proceedings of the Workshop on Figurative Language Processing*, pages 7–16, New Orleans, Louisiana. Association for Computational Linguistics.

Nick Stanko. 2001. Use of language in diplomacy - diplo resource. In *Langauge and Diplomacy*.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. 54(1):123–154.

Karolina Zaczynska, Peter Bourgonje, and Manfred Stede. 2024. How diplomats dispute: The UN security council conflict corpus. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Džemal Špago. 2020. Rhetorical questions as aggressive, friendly or sarcastic/ironical questions with imposed answers. *ExELL*, 8(1):68–82.

## A Appendix Annotation Guidelines Extension

The following text is taken from the annotation guidelines and explains the annotations for Figurative Language. Figure 3 shows the annotation steps for Conflict types with the refined annotation guidelines.

Based on the results of our UNSC expert annotation experiments, we have expand the annotations guidelines by (Zaczynska et al., 2024) by including a new Conflict type, FIGURATIVE LANGUAGE (FIGL), which includes sarcastic statements (label: SARCASM) or rhetorical questions (label: RHETORICAL QUESTION) that serve to express a negative evaluation of another country. Sarcasm and rhetorical questions are figurative language, meaning they convey a message that is different from what is literally said (Skalicky and Crossley, 2018; Ducret et al., 2020).

**Sarcasm.** Sarcasm is defined as specific instances of verbal irony which serve to provide ironic criticism or praise that is somehow contrary to reality (Skalicky and Crossley, 2018). Sarcastic sentences are likely to be semantically or emotionally incongruent with their preceding sentences but also incongruent with the situation in which sarcasm is used. Detecting sarcasm might not be straightforward when only looking at the text. Thus,

the annotators must also rely on understanding of the context beyond the statement to discern between sarcasm and sincerity. Following Moreno-Ortiz and García-Gámez (2022); Joshi et al. (2017) we annotate sarcasm as negative in nature, and the message must contain some form of criticism and an implied negative sentiment for it to be classified as Conflict type SARCASM.

**Rhetorical Questions.** A rhetorical question is an utterance that has the structure of a question does not expect an answer (Rohde, 2006). It can be seen as a mechanism to express sarcasm (Moreno-Ortiz and García-Gámez, 2022). Rhetorical questions are often lexically and syntactically not easily distinguishable from other types of questions. However, there are some linguistic cues that make a question more obviously rhetorical: Does it include strong negative polarity items (*at all*, *any*, *ever*)? Can it be preceded by the expression after all and followed by a *yet*-clause (Špago, 2020; Comrie and Sadock, 1974)?

In summary, the annotators mark EDUs as FIGURATIVE LANGUAGE if the following applies: Does the EDU/sentence use irony that indicates a negative evaluation or critique toward a country? This can be signified by: 1) SARCASM, meaning that the text expresses an evaluation whose literal polarity is the opposite of the intended polarity, or 2) RHETORICAL QUESTION, which is asked not primarily to elicit information, but to make a (negative) statement.

## B Prompt Used for Lexical Marker Extraction

The following shows the prompt we used to extract the lexical markers and the categories per EDU from or corpus.

**System / Instruction to the Model**
You are an expert language processing system. Please analyse the text below for verbal conflicts or critique.

—

### Task

Given the following text:
{{TEXT_EDU}}
Perform **three** steps:

1. **Check for Presence of Lexical Markers**
   Determine whether the text contains any words/phrases that indicate negative evaluations, which we define as critique or distancing from another entity (person, country, group, etc.). Specifically, look for any of the following:

   - "Adjectival_Attribution": Adjectival attributions (e.g., *bad*, *dreadful*, *worrying*)
   - "Noun": Nouns with a negative connotation (e.g., *traitor*, *annexation*)
   - "Adverb": Adverbs that intensify criticism (e.g., *poorly*, *even*, *only*)
   - "Verb": Verbs with a negative connotation (e.g., *infiltrating*, *invading*)
   - "Negation_Phrase_or_Quantifier": Negation phrases and quantifiers (e.g., *not at all*, *not a single*)
   - "Evaluative_Pattern": Recognisable evaluative patterns (e.g., *It is unfortunate that…*, *There is something worrying about…*)
   - "Instructive_Words": Strong instructive words (e.g., *urge*, *must*, *warn*, *demand*)
   - "Emotive_Words": Strong emotive words (e.g., *condemned*, *armed*, *shocked*)

   **Response**: Indicate **Yes** or **No** (e.g., ‘Present?: Yes‘ / ‘Present?: No‘).

2. **Extract Lexical Marker Categories**
   If you found negative markers, list which categories these markers belong to (e.g., "Adjectival_Attribution",” "Negative_Noun", "Negation_Phrase_or_Quantifier", etc.).
   *Response**: Provide the categories as a comma-separated list, choosing from the following categories: ’Adjectival_Attribution’, ’Noun’, ’Adverb’, ’Verb’, ’Negation_Phrase_or_Quantifier’, ’Evaluative_Pattern’, ’Instructive_Words’, ’Emotive_Words’ or write ‘None‘ if no markers are found.

3. **List the Lexical Markers**
   List the actual words or phrases that caused you to identify negative evaluations. **Response**: Provide a comma-separated list of markers (e.g., ‘bad, dreadful, invaded‘), or write ‘None‘ if no markers are found.

—

### Output Format

- Present?: [Yes or No] - Lexical Marker Categories: [comma-separated categories or 'None'] - Lexical Markers: [comma-separated words/phrases or 'None']

## C Most Frequent Lexical Marker of Negative Evaluation

| LM Category | 10 most frequent words |
|---|---|
| **Noun** | crisis (45), violence (33), terrorists (31), war (30), threat (26), conflict (21), terrorism (20), weapon (18), armed (18), crime (17) |
| **Instructive Words** | must (100), urge (17), call (10), should (8), demand (6), reject (3), halt (2), strongly (2), condemn (2), immediate (2) |
| **Adjectival Attribution** | illegal (19), serious (17), difficult (10), unacceptable (10), illegally (7), arm (6), dangerous (6), critical (6), criminal (6), deeply (5) |
| **Negation Phrase or Quantifier** | not (99), no (59), can (23), without (22), do (19), nothing (14), never (8), despite (6), non (3), nor (3) |
| **Verb** | destabilize (19), condemn (17), attack (15), undermine (14), threaten (13), kill (13), seize (12), shoot (12), destroy (10), fail (9) |

Table 3: Most frequent Lexical Markers (LM) found per category, lemmatised using SpaCy library (model *en_web_core_sm*).

## D Flowchart Conflict Annotations

## E Visualisation Streams of Conflicts between Source and Target Comparing both Corpus Versions

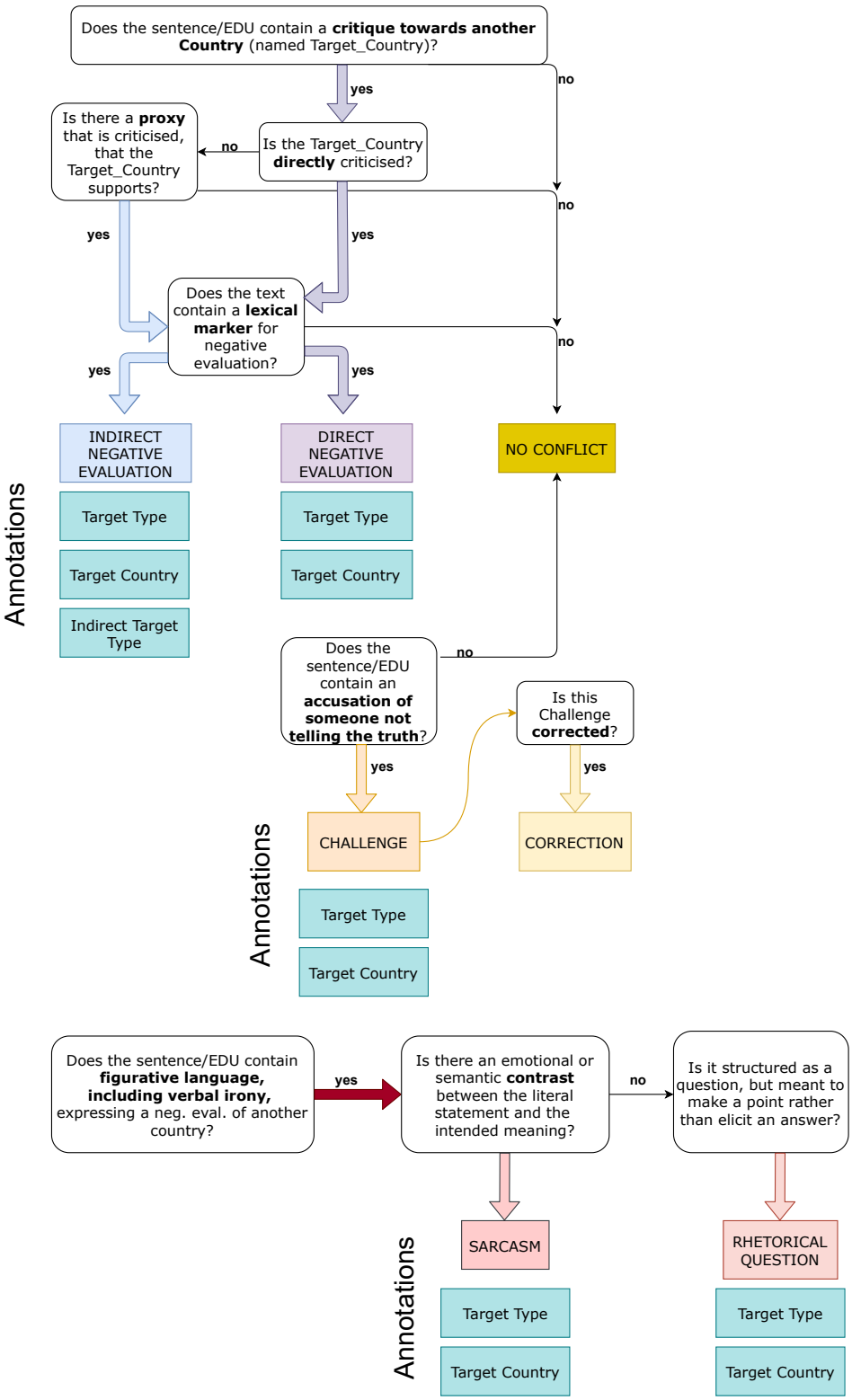


Figure 3: Annotation Steps of Conflict Type and Target Annotations Visualised in a Flowchart.

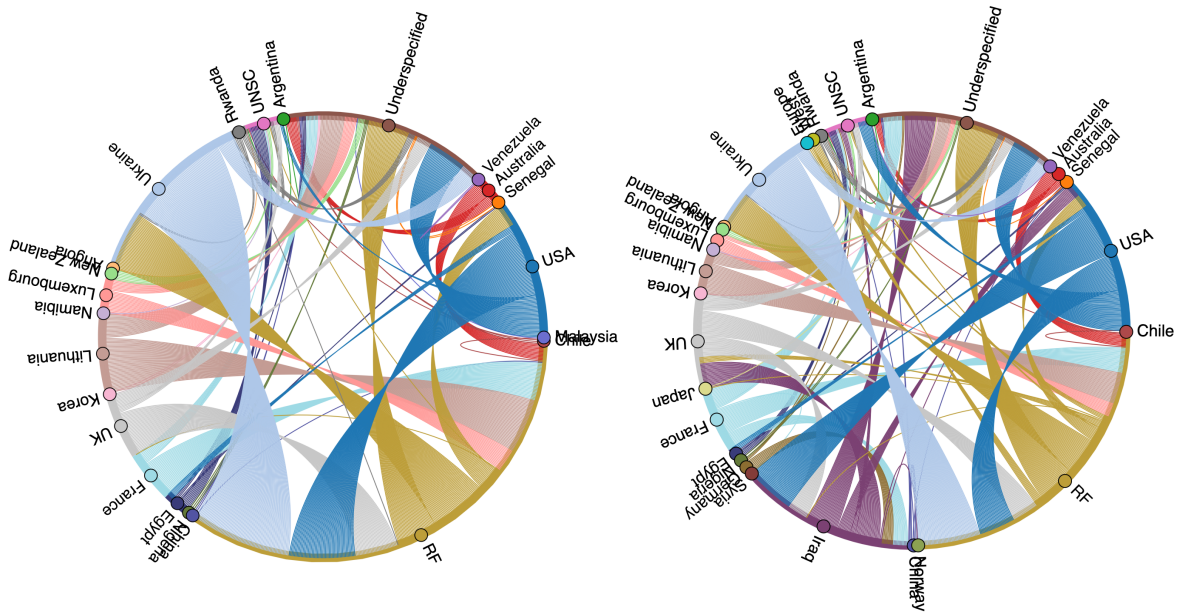


Figure 4: Visualisations of the source and target of Conflicts from the original UNSCon (left) and the extended UNSCon (right circle). An HTML version of the figure is available in our GitHub repository. RF stands for the Russian Federation, UK for the United Kingdom of Great Britain and Northern Ireland, and USA for the United States of America.

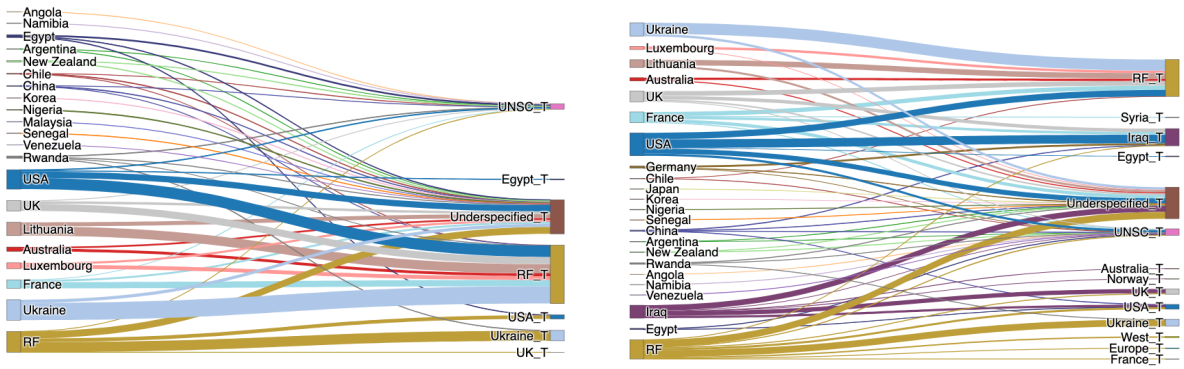


Figure 5: Sankey graphs of the source and target of Conflicts from the original UNSCon (left) and the extended UNSCon (right sankey). The source is on the left side, the target (marked by _T) is on the right side.