How Diplomats Dispute: The UN Security Council Conflict Corpus

Karolina Zaczynska, Peter Bourgonje, Manfred Stede
University of Potsdam, Applied Computational Linguistics
Potsdam, Germany
{lastname}@uni-potsdam.de

Abstract

We investigate disputes in the United Nations Security Council (UNSC) by studying the linguistic means of expressing conflicts. As a result, we present the UNSC Conflict Corpus (UNSCon), a collection of 87 UNSC speeches that are annotated for conflicts. We explain and motivate our annotation scheme and report on a series of experiments for automatic conflict classification. Further, we demonstrate the difficulty when dealing with diplomatic language – which is highly complex and often implicit along various dimensions – by providing corpus examples, readability scores, and classification results.

Keywords: Sentiment Analysis, Diplomatic speech, Conflict Detection

1. Introduction

The UN Security Council (UNSC) is a unique institution that debates security challenges in changing geopolitical constellations. Major multilateral agendas are discussed, in language that is usually highly diplomatic and restrained. Until now, there has been little work on how to formalize conflicts (i.e., disputes or disagreements) in a diplomatic setting. Most approaches to the established task of *disagreement detection* are based on direct responses to statements in (spontaneous) dialogs (De Kock and Vlachos, 2022; Zhang et al., 2018; Chang and Danescu-Niculescu-Mizil, 2019). However, the debates in the UNSC are – with some exceptions – accurately prepared speeches of several minutes.

More importantly, when criticizing someone in the Council, diplomats often avoid direct verbal attacks, in line with diplomatic practice of face keeping and negotiation (Nair, 2019). Instead, they express criticism rather implicitly, sometimes without clear mention of the member that is being criticized. Linguistically, this is done through the use of third person, passive or other indirect constructions. In the following example, discussing the situation in Crimea right after the annexation in 2014, the speaker uses the adjective “external”, but inferring which member or country is being held responsible for the criticized situation requires at least the previous sentences, and probably the entire political discourse.

(1) Again, the external anti-Ukranian and anti-Western propaganda machine is in full swing, inciting suspicion, mistrust and hatred waiting to explode.
(UNSC_2014_SPV.7154, Lithuania)[1]

This highly implicit style renders the detection of disagreements, or conflicts, as we will continue to call them, a very difficult task. To the best of our knowledge, neither the challenge of implicit reference has been addressed, nor is there a detailed characterization of conflicts in the Security Council in the first place. We believe that understanding the nature of conflict in UNSC debates is important, because the topics being discussed in the Council, and the way different countries respond to them linguistically, can be a precursor to their next actions, which are often of global significance.

This work presents a new dataset with annotations for conflicts in UNSC debates, where the definition of conflicts is specifically tailored to diplomatic language. Our main contributions are (1) the formal characterization of conflict in UNSC debates, resulting in annotation guidelines, (2), the annotation of 87 speeches following these guidelines, and (3) first classification experiments on the resulting annotations. We first discuss the theoretical background (Section 2). Subsequently, we describe the annotation guidelines (Section 3) and the resulting corpus (Section 4). We experiment with classification models for sentiment detection, and we fine-tune different BERT-based models[2] (Section 5) and discuss their results (Section 6). Finally, we sum up the main conclusions and discuss planned future work (Section 7).

2. Background and Related Work

Here we discuss the nature of UNSC debates and explain how we arrived at our definition of conflicts.

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

[2] The repository with the guidelines, dataset, and code is available at: `https://github.com/linatal/UNSCon`

LREC-COLING 2024, pages 8173–8183

2.1. UNSC Speeches

Speeches given at the UNSC can contain formal reports and also opinions (Kutateladze, 2020). Each debate has pre-defined agenda items. The speeches are usually pre-written, precisely formulated, and several minutes long (our selected speeches have an average of 28 sentences).[3] The speeches usually consist of two main parts, beginning with a preamble in which the president and external invited speakers are thanked. The body of the speech consists of reports or opinions, where countries explain their stance or commitment towards an action. The tone of voice is typically formal, words are selected with great caution and the style is usually more or less neutral (Kutateladze, 2020). One of the goals of diplomacy is keeping the balances of interests and political compromise between different parties (Slavic and Kurbalija, 2001), and linguistic vagueness is one technique for widening the degree of applicability and acceptance of policies (Scotto di Carlo, 2012). We note, however, that particularly the Crimea agenda is somewhat atypical in its more confrontational and slightly less diplomatic style. From all speeches collected by Schönfeld et al. (2019), we take 62 from the Ukraine/Crimea agenda (henceforth "Ukraine"), and 25 from the Women, Peace and Security (henceforth "WPS") agenda (see Section 4). These 87 speeches in total are what the statistics and experiments in the rest of this paper are based on. There is a body of work based on the existing corpus coming from computational linguistics and political science. To name just a few, some papers deal with the extraction of country mentions in the in the UNSC using Wikidata for Named Entity Linking (Glaser et al., 2022) and Named Entity Recognition (Ghawi and Pfeffer, 2022), others work with network analyses on UNSC topics from Afghanistan debates (Steffen Eckhard and van Meegdenburg, 2021) and on discourse in climate change (Scartozzi, 2022).

2.2. Linguistic Complexity

One challenge in dealing with diplomatic debates is their higher linguistic complexity than an average wikipedia or newspaper article. We compare our corpus to a set of one million sentences from Wikipedia[4] and 2.200 news articles from the Wall Street Journal (using the Penn Discourse Treebank 3.0 (Prasad et al., 2019)) (see table 1).

To assess complexity, we calculate readability scores.[5] Particularly, we use (1) the Gunning Fog index, which considers sentence length and the number of words longer than three syllables; and (2) the Automated Readability Index (ARI), which essentially counts the number of characters per word. For both scores, the UNSCon has the highest scores, which indicates the use of unusual and long words and relatively long sentences compared to, for example, the ones in Wikipedia articles.

Because the metrics are mainly word-based and do not say much about grammatical complexity, we also use dependency tree[6] depth as a proxy for linguistic complexity (inspired by Oya (2011); Pitler and Nenkova (2008); Schwarm and Ostendorf (2005)). For every sentence, we extract the maximum tree depth (the token with the longest path to the root node). We average this number for all sentences in the corpus to arrive at a measure for average tree depth of the corpus. Highly embedded structures will have a high depth, indicating higher grammatical complexity of the sentences. The results show that the sentences in our corpus have the highest score with 6,74, followed by PDTB-3 with 6,11 and Wikipedia articles with 5,42. The dependency trees also show that 22% of the sentences' speeches in the UNSCon contain passive constructions. These are often used to obfuscate or leave implicit the agent in a sentence, which is particularly challenging for our annotation task.

2.3. Definitions

We define a conflict as an expression of critique or distancing from the positions or actions of another country present at the Council during the debate. Following Gleditsch (2020); Deutschmann et al. (2020); Maerz and Puschmann (2020), we are interested solely in the linguistic expression of conflict, regardless of whether there are underlying violent or military confrontations. In this view, a conflict consists of a target, which is the entity being evaluated, and a negative evaluation (NegE) of that target. The holder of the evaluation is always the country/member state represented by the speaker. We exclude *reported* conflicts that exist between actors not including the speaker.

Conflicts partially overlap with what in the literature (Galley et al., 2004; Allen et al., 2014; Chang and Danescu-Niculescu-Mizil, 2019; Somasundaran et al., 2007) is often referred to as *disagreement*. The nature of our domain though, in which diplomats often know the position of other countries in advance (and there is no turn-taking as in ordinary conversation) renders the literature's definition (which is much more based on direct,

[3] Sometimes diplomats decide to comment on what has been said earlier in the debate in an often much shorter and rather spontaneous second speech at the end of the debate.

[4] Collected in December 2020.

[5] `https://github.com/cdimascio/py-readability-metrics`

[6] Obtained using spaCy's `en_core_web_sm` model.

Dataset	sents./ speech or article	tokens/ sents.	ARI	Gunning Fog	depend. tree depth	passive / sents.
UNSCon	28,33	25,57	14,69	16,47	6,74	0,22
PDTB-3	20,41	22,23	12,86	14,05	6,11	0,2
Wiki -	-	20,06	10,1	12,56	5,42	0,29

Table 1: Comparing UNSCon with WSJ articles (PDTB-3) and sentences from English Wikipedia looking at dataset statistics, readability scores and sentence complexity.

interactive dialog situations) only partially relevant.

Furthermore, our task has partial overlap with the more popular NLP task of sentiment analysis, or stance detection, but differs in that we focus on only the negative half of evaluative statements.

We aim to annotate utterances in transcriptions of speeches, so the conflicts have to be present and salient enough, with some linguistic marking present (i.e., the conflict cannot be non-verbal, only inferrable from political discourse or events not discussed in the speech itself). We based our definition of evaluation on Martin and White (2007) in terms of positive and negative stance (valence). In addition, Conrad and Biber (2000) and Taboada (2016) provide helpful contributions on the usage of adjectives and adverbials to express an attitude toward an entity. In the next example, adverbial constructions are used to express a NegE of a member of the Ukrainian parliament (referenced through the possessive pronoun *her*) and of groups in Ukrainian politics. Lexical markers expressing NegE are in bold:

(2) We can only imagine the thoughts that must be churning in the minds of her **brutal** fellow partisans. And that is not even the most **radical** group on the Ukrainian political spectrum. (UNSC_2014_SPV.7154, Russian Federation)

The following example includes an evaluative noun and evaluations expressed at phrasal level:

(3) Ukraine's **traitor** Yanukovich, who **abandoned his country and fled**, **opening the floodgates** to Crimea's annexation, is being pushed again into the daylight to clear the way for Ukraine's further dismemberment. (UNSC_2014_SPV.7154, Lithuania)

Besides NegE, we annotate statements that evaluate the perceived truthfulness (or lack thereof) of another statement. We call these challenging and correcting statements (CC). In the next section, NegE and CCs are described in more detail.

3. Annotation Guidelines

The discussion in the previous section leads us to define guidelines for annotating conflicts in our data. Conflicts can be expressed by directly criticizing another country (*Direct NegE*) or by indirectly addressing the critique to a surrogate entity that serves as a proxy (*Indirect NegE*) (an example is included in Section 3.3). Both conflict types need a lexical marker of NegE. In addition, we look at *Challenging* statements, accusing the target of not telling the truth and the *Correction* of that allegedly false statement. Figure 1 graphically illustrates the categories we set out to annotate.

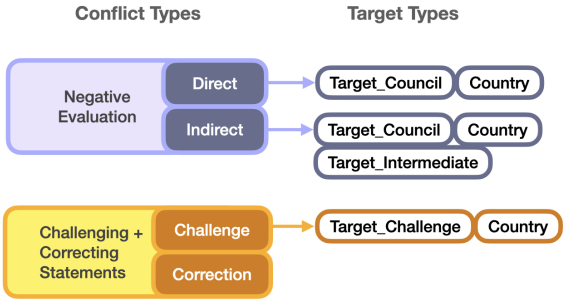

Figure 1: Conflict and Target Types

3.1. Units of Annotations

As explained in Section 2.3, the linguistic markers of evaluation are hard to restrict to some syntactic category, and often challenging to even pinpoint to specific words. To make our annotations more robust and consistent, we therefore need a larger unit. We found the sentence-level to be too coarse though, as preliminary annotation experiments sometimes resulted in single sentences ending up with multiple, conflicting annotations. To find the right granularity, we borrow the notion of Elementary Discourse Units (EDUs) from Rhetorical Structure Theory (Mann and Thompson, 1988). EDUs are usually sentences or certain types of clauses; see Stede et al. (2017) for more details. We manually segmented the corpus into EDUs, and applied labels based on EDU spans.

3.2. Direct NegE and Council Targets

We define Direct NegE as conflicts where the NegE is overtly directed at someone present at the Council. The addressee of the Direct NegE is called

Target_Council and is annotated, too. We define six possible Target_Council types:

1) A previous or upcoming **speaker** or **speech** ("In her last speech we heard Mrs. . . . ");
2) A **Country**, including governments or representatives of a country ("Frau Merkel" for Germany, "Verkhovna Rada" for the Ukrainian Parliament):
3) A **Group of Countries** ("the African Union", "permanent members of the Council");
4) The **UNSC** ("the Council"), often referred to via self-referencing ("we");
5) **Self-targeting**: Diplomats may refer to themselves or the country they are representing using pronouns ("We" or "I");
6) **Underspecified**: For vague mentions such as "some in this chamber" or "the international community", but where the annotator feels that it is implicitly clear who is meant; otherwise there is no annotation.

3.3. Indirect NegE and Intermediate Targets

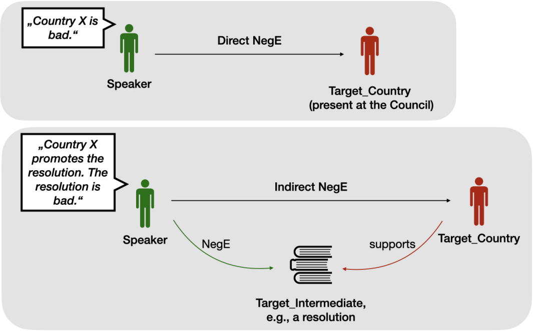

Figure 2: Schema of Direct and Indirect NegE

Indirect NegE are conflicts where some intermediate is criticized instead of the Target_Council directly. In this case, we annotated a Target_Intermediate, in addition to the Target_Council, to whom the criticism is actually directed. We define five possible *Target_Intermediate* labels:

1) A **Policy or Law** (e.g., a resolution);
2) A **Person** which is not an official UNSC representative of a Country or Group;
3) A **UN-Organization** other than the Council;
4) A **Non-Governmental Group** ("the terrorist groups", "separatists") supposedly working for a Country;
5) The **other** label can be applied to intermediate targets that fit none of the above.

The following example includes an Indirect NegE:

(4) That general context is important to understanding our view of the draft resolution submitted by the United States (S/2014/189) for the Council's vote today. [...] – *no Conflict*

We can not go along with its basic assumption, which is to declare illegal the planned referendum of 16 March [...]. – *Indirect NegE, Target_Intermediate: Law/Policy, Target_Council: USA* (UNSC_2014_SPV.7138, Russian Federation)

Here the draft resolution (the Target_Intermediate) is criticized, with the Target_Council being the United States, who proposed the resolution. This schema is illustrated in Figure 2.

3.4. Challenging and Correcting Statements (CC)

Statements that attack an opponent's claim as untrue are called *Challenges*. We annotate EDUs that report allegedly untrue information, and the *Correction* of that statement, describing the truth as perceived by the speaker. Challenging statements evaluate the opponent for truthfulness but do not necessarily need linguistic markers of NegE. The following is an example for a CC:

(5) To conclude, one of our colleagues said that Kyiv had extended a hand to Moscow and that we had refused to reciprocate. – *Challenge, Target_Challenge: Underspecified*

But the problem is not with Moscow; it has to do with the fact that Kyiv should have been the one to extend a hand to its people and regions [...]. – *Correction*

(UNSC_2014_SPV.7138, Russian Federation)

The Target for Challenging statements (*Target_Challenge*) is always a state or representative present at the Council, and can have the same six target types as Target_Council. We do not annotate Targets for Corrections. The annotation is not to be confused with fact-checking, or assessing the validity of a claim based on evidence. Our focus is solely on the speaker's attitude toward something said, not on verifying this against objective truth.

3.5. Country

Next to target types, we asked the annotators to identify a country for the Target_Council or Target_Challenge where possible. Only a country present at the debate, not groups of countries ("the West", "the Council" etc.) are annotated.

4. Dataset

Here we describe the corpus selection and annotation procedure, and we present corpus statistics.

4.1. UNSCon: Selecting the Speeches

The UNSCon is grounded in the UNSC corpus by Schönfeld et al. (2019). The transcripts are all in English. We selected two topics with different expected potential for conflicts. The first agenda (henceforth "Ukraine") is the Ukraine conflict in 2014 after the annexation of Crimea (and before the Minsk II agreement). The second agenda (henceforth "WPS") is the Women, Peace and Security agenda. We expect the first agenda to contain more intensified and direct expressions of conflict.[7] For both topics we selected debates such that they further maximize the probability of finding expressions of conflict in the speeches, by choosing debates (1) dealing with resolutions with an unanimous voting, (2) with an average sentiment score below 0, using Lexicoder (Young and Soroka, 2012) (see also Section 5.1) and (3) based on review of the discussed sub-topics together with a political scientist. We discarded from this preselection any speeches given by external experts, and included speeches from permanent members of the UNSC[8], and from countries having more than one contribution to the debate. This left us with 87 speeches from 6 debates (two from WPS, four from Ukraine). We automatically delete line breaks that still originate from the PDF conversion to plain text from the original corpus, tokenized the text using spaCy[9], and excluded preamble statements by the president.[10] The speeches were then segmented into EDUs manually. The UNSCon has a total of 4.726 EDUs and 2.437 sentences. Each debate has 14,5 speeches on average and each speech has 28 sentences and 764 tokens on average. For more statistics we refer to Table 2.

4.2. Annotation Procedure

The annotations were done by two computational linguistics students who are familiar with the UNSC data from a former project. Our guidelines were iteratively refined over the course of half a year; final annotations were conducted within a month. We used the INCEpTION tool.[11] In the final annotation phase, we had weekly meetings to discuss challenging cases. The annotators were given the list of names and countries present at the debate, and a summary of a debate from a website[12] providing reports on the Council's activities. The speeches were presented to the annotator in the order they were delivered during the debate, and the representative's country was visible.

Debate	#spch orig.	#spch UN-SCon	#sents UN-SCon	#tok UN-SCon
Ukraine				
SPV.7138	21	17	345	8.797
SPV.7154	28	17	454	11.264
SPV.7165	22	13	501	12.270
SPV.7219	36	15	511	12.022
sum Ukr.	107	62	1.811	44.353
WPS				
SPV.7643	18	16	231	6.993
SPV.7658	75	9	395	10.985
sum WPS	93	25	626	17.978
Sum	200	87	2.437	62.331

Table 2: Overview of UNSCon

To calculate inter-annotator agreement (IAA), 30 percent of the dataset was annotated by both annotators after the consultation phase. For NegE we report a Cohen's Kappa of 0,72, i.e., substantial agreement according to Landis and Koch (1977). Additionally, we calculated Krippendorff's Alpha for all values, as it allows multi-label annotations and takes into account partially-overlapping annotations. For NegE (two labels) we have an agreement of 0,64, for Target_Council (seven labels) 0,70, for Country NegE (five labels) 0,70 and for Target_Intermediate (six labels) 0,64. For Challenge Type and the Targets we report moderate agreement of 0,54 (with two, seven, or five different labels). We discuss some of the challenges the annotators faced in Section 6.3. In earlier experiments, we asked two UNSC experts to annotate sentences using a broader definition of conflict, defined as "a collision of competitive positions verbalized in political speeches where the target must be present at the Council." They annotated 17 speeches from one of the Ukraine debates (UNSC_SPV.7154). These annotations had an IAA of 0,29 Cohen's Kappa, showing that even experts often disagree on what a conflict statement is, highlighting the importance of an operationalizable definition of conflict in diplomatic language, which we consider one of the main contributions of this paper.

4.3. Distribution of Labels

The EDUs are labelled as Conflict (NegE or CC) or left unlabelled ("No Conflict"). The annotations have multiple layers of labels, and each layer has two (NegE and CC), six (Target_Intermediate) or up

[7] Note that (Schönfeld et al., 2019) includes debates until 2019, so more recent developments in Ukraine are not included.

[8] The permanent members are: China, France, Russian Federation, United Kingdom, and United States.

[9] https://spacy.io/

[10] Presidential statements mainly serve to organize the debate in accordance with the rules of procedure, including stating the speaker order, welcoming the participants and announcing upcoming speakers.

[11] https://inception-project.github.io/

[12] https://www.securitycouncilreport.org/

to seven (Target_Council and Target_Challenge) labels. The Target_Country layer potentially has as many labels as states in the Council, but in our corpus, the annotators used only five unique countries as targets of a conflict. See Table 3 for details. The annotations were curated and disagreements resolved by one of the authors of this paper. We have seven different label types: Two conflict types (NegE and CC) and five different target types (Target_Council, Target_Challenge, Target_Intermediate, Country for NegE, Country for CC). From all 4.726 EDUs, 1.501 have a conflict annotation. 28,79% of all EDUs in Ukraine debates are NegE, compared to 12,16% in WPS debates. There are fewer Challenge/Correction annotations, with 5,64% for Ukraine and 0,71% for WPS.

In Figure 3 we look at the distribution of Conflict Types for the topics *Ukraine* and *WPS* in the UNSCon in more detail. While the WPS-subcorpus includes less EDUs marked as Conflict than the Ukraine-subcorpus, in both the distribution of Direct and Indirect NegEs is similar (Ukraine: 51% *Direct NegE* and 33% *Indirect NegE* for Ukraine, 86% and 14% for WPS respectively). We want to point out that the frequency for Challenging and Correcting EDUs is much higher for Ukraine than for WPS. Challenges are accusing someone else of lying, which is a rather intense critique, and the increased frequency in the Ukraine debates is perhaps unsurprising, given its confrontational character. Many of the found Conflicts (marked as Indirect NegE) are not naming a target explicitly, which is in accordance with what we found in the literature about diplomatic speech (Kutateladze, 2020; Slavic and Kurbalija, 2001). We were surprised that although we bound our Conflicts by several rules, such as requiring an explicit lexical marker, we see a lot of directly expressed critique, which is not what is suggested by other literature that emphasizes the implicit nature of the language used. We assume that, at least for the Ukraine speeches, this is due to the topic, and that other debates will most probably have a lower density of Conflicts.

Looking at the Targets, we see that for the Ukraine debates, the majority of EDUs annotated as NegE specify a *Country* that is directly addressed (76,08%). The second-most frequent target is *Underspecified* (21,93%, 243 EDUs), implying that a Target_Council can be inferred, but is not explicitly mentioned. In the WPS debates, the speeches seem to avoid direct addressing, with the majority of Target_Councils being *Underspecified* (48,17%, 79 EDUs). The second-most frequent target is the Council (label: *UNSC*) (34,76%, 57 EDUs) and only 12,2% (20 EDUs) directly target a Country.

Labels for NegE	**#EDU**	**%**
Indirect NegE	501	10,6
Direct NegE	771	16,31
L. for Target_Council	**#**	**%**
Countries_Group	12	0,25
Country	861	18,22
Self-targeting	3	0,06
Speaker_Speech	1	0,02
UNSC	71	1,5
Underspecified	322	6,81
-NONE-	2	0,04
L. for Target_Intermediate	**#**	**%**
Law_Policy	142	3,0
Non-Governm_Grp	295	6,24
Other	35	0,74
Person	31	0,66
UN-Organization	7	0,15
-NONE-	762	16,12
L. for Target_Country (for NegE)	**#**	**%**
Egypt	14	0,3
Russian Fed.	705	0,15
Ukraine	133	2,81
USA	47	0,99
-NONE-	373	7,89
No NegE	3454	73,09

Labels for CC	**#**	**%**
Challenge	101	2,25
Correction	128	2,84
L. for Target_Challenge	**#**	**%**
Countries_Group	4	0,08
Country	79	1,67
Self-targeting	0	0
Speaker_Speech	2	0,04
UNSC	0	0
Underspecified	16	0,34
-NONE-	128	2,71
L. for Target_Country (for CC)	**#**	**%**
Russian Fed.	66	1,47
Ukraine	5	1,11
United Kingdom	2	0,04
USA	8	0,18
-NONE-	148	3,29
No CC	4497	95,15

Table 3: Distribution of labels per EDU in UNSCon

5. Classification Experiments

5.1. Lexicon-based Sentiment Classifier

Since (negative) evaluation partially overlaps with the NLP task of sentiment analysis, we use Lexicoder (Young and Soroka, 2012), a sentiment classifier that was tailored to political texts and is popular in Social Science research. It counts words having positive, negative, negated-positive or negated-negative sentiment in a dedicated dic-

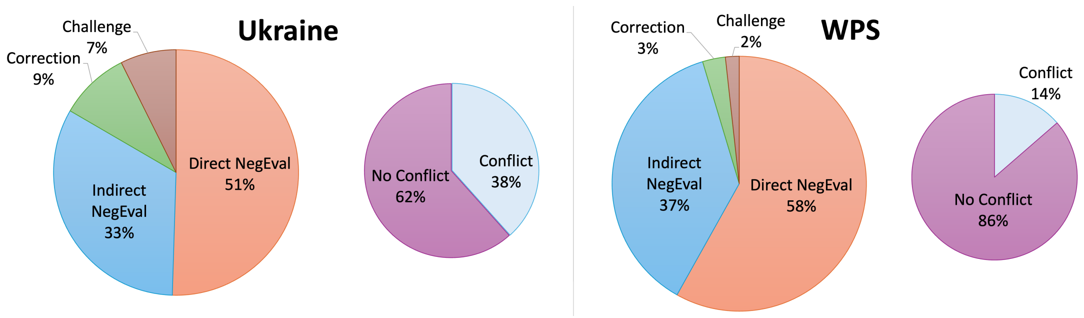

Figure 3: Distribution *Conflicts* and *No Conflicts*, and distribution of Conflict Types per EDU for two subcorpora *Ukraine* and *WPS* in UNSCon

tionary, and calculates a sentiment score from this. As a small modification, we remove the entries *unite** and *united** from the dictionary since mentions of the United Nations would otherwise inappropriately influence the score. Since the annotations are sentence-based for most cases, and the lexicon based approach is taking the average of all found positive or negative words, we decided to map our EDU annotations to their host sentences. This means that for some cases (where multiple EDUs within a sentence had different labels), the sentence-based annotations are not 100% accurate. Conflicting annotations are resolved by using the last one (i.e., the annotation relating to the token (span) appearing later in the sentence). We map our Conflict categories to binary values (one for Conflict, zero for none). Lexicoder produces a score between -1 and +1; for our experiments, everything below 0 is seen as conflict.

5.2. BERT

To test pre-trained language models on our task, we deploy three different BERT-based models. First, we fine-tune `distilbert-base-uncased` (Sanh et al., 2019) on our data. This is a distilled version of the BERT base model trained on the BooksCorpus (Zhu et al., 2015) and English Wikipedia. We compare this to a model trained on the task of sentiment analysis `sentiment-roberta-large-english` (Hartmann et al., 2023), and to a model trained on debates and arguments `roberta-argument`[13]. We train models for binary classification of conflict, and for differentiating between NegE, CC and no conflict. Because we have a relatively small number of data points, we set learning rate to 1e-5, and we use a batch size of 32, with 2 training epochs and a weight decay of 0.01. We train the classifier to assign labels for EDUs. All scores reported for the BERT-based models are the result of 10-fold cross-validation.

6. Results and Discussions

The following sections present the classification results and discuss the main challenges the annotators faced during annotation.

6.1. Lexicoder Results

Approaching our conflict annotation task as if it were sentiment analysis, we achieve a weighted f$_1$-score of 0,71, and a macro f$_1$-score of 0,65. In addition to binary classification for the entire corpus, we look at only those sentences marked as NegE, as these need a lexical marker (as opposed to CC annotations). We expect Lexicoder to perform better on this subset, but find that weighted f$_1$-score improves only slightly, to 0,72 (+0,02) and 0,65 macro. Despite the relatively good scores, we clearly see the shortcomings of a lexicon-based approach when analysing the results. The following example shows how some words that have a positive value in the lexicon ("colleagues" as part of the arguably negatively connoted "western colleagues"), are actually used in a negative way by the speech. While this might be a matter of lexicon adjustment, for many other cases (like "support" in the example), interpretation highly depends on context, for which a purely lexicon-based system does not suffice.

(6) However, the Kyiv authorities have chosen the *wrong*(-) path at every turn, with the *support*(+) of its Western *colleagues*(+). – *Direct NegE, Lexicoder: pos.* (UNSC_2014_SPV.7219, Russian Federation)

6.2. BERT Results

For the binary classification task, for `distilbert-base-uncased` we achieve a weighted f$_1$-score of 0,71 and a macro f$_1$-score of 0,65, which is comparable to Lexicoder performance. Note that scores cannot be compared directly though, since Lexicoder experiments were sentence-based, whereas BERT-based experiments are EDU-based. We

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

note that a majority vote classifier scores 0,55 and 0,41 for weighted f_1-score and macro f_1-score, respectively. For `sentiment-roberta-large-english`, scores are considerably better, with 0,76 and 0,72 for weighted f_1-score and macro f_1-score, respectively. The best performance is achieved by `roberta-argument`, with 0,78 and 0,74, respectively. When training a classifier for the three labels of NegE, CC or *No Conflict*, we see a performance drop, with weighted f_1-scores for the three models being 0,63, 0,73 and 0,74, and macro f_1-scores 0,36, 0,47 and 0,48 (with majority vote weighted and macro being 0,63 and 0,27, respectively).

This might be due to the size and imbalanced distribution of the classes. The number of instances for the three-class setup are 3.231 (No Conflict), 1.274 (Negative Evaluation) and 225 (ChallengeCorrection). In future work, we plan to include sampling methods for the classification experiments and expand the dataset. Overall, these results indicate that training for sentiment analysis helps, but specifically training on debates and arguments results in the largest gain. Compared to many other NLP tasks though, we consider f_1-scores to be low, stressing the difficulty of dealing with conflicts in diplomatic language.

	$\text{BERT}^{\text{distil}}_{\text{uncased}}$	$\text{sRoBERTa}^{\text{large}}_{\text{english}}$	$\text{RoBERTa}^{\text{argument}}$
Conflict / No Conflict			
precision	0.73	0.76	0.78
recall	0.74	0.77	0.78
f_1-weighted	0.71	0.76	**0.78**
f_1-macro	0.65	0.72	**0.74**
accuracy	0.74	0.77	**0.78**
maj. voting	0.55	0.55	0.55
NegE / CC / No Conflict			
precision	0.65	0.71	0.72
recall	0.71	0.75	0.76
f_1-weighted	0.73	0.72	**0.74**
f_1-macro	0.47	0.47	**0.48**
accuracy	0.71	0.75	**0.76**
maj. voting	0.62	0.62	0.62

Table 4: Classification results of 10-cross validation over 4.726 EDUs for fine-tuned pre-trained models. Precision and recall relate to the weighted-average. sRoBERTa stands for sentiment-RoBERTa.

6.3. Challenges for Annotators

In the following, we discuss the main sources of disagreement that we observed in the annotations.

6.3.1. Naming the Target

Annotating the target of the conflict is crucial, since it defines a conflict's status and type. This is far less trivial than it may sound. One reason is that for Direct NegE, the Target_Council can be referred to in different ways:

- Name of the speaker (e.g., *Mr. Smith*);
- Name of the country or group of countries (e.g., *the African Union*, *the Council*);
- Using different aliases referring to a government (e.g., *Berlin* instead of *Germany*) or through pejorative names (e.g., *the Kyiv regime*);
- Using self-referencing formulations (*we*).

Self-references are particularly interesting since in a conflict, one usually accuses the other. Sometimes self-referencing is used though, to indirectly and diplomatically target others in the Council, as in the following example:

(7) Please, let us refrain from any accusations or speculation as to why Russia is trying to do what it is doing. (UNSC_2014_SPV.7154, Russian Federation)

6.3.2. Council and Intermediate Target

For some cases, it was difficult to distinguish between a Direct or Indirect NegE. Consider the following examples:

(8) According to Turchynov, the people of south-eastern Ukraine must end their protests by the morning of Monday, 14 April, lest armed force be used. *– no Conflict*

However, **the protesters' interests and opinions have not been taken into account or even discussed**. *– Indirect NegE, Target_Council: Country, Ukraine, Target_Intermediate: Non-Governmental Group*

As a result, **blood has already been shed** in the South-East and the situation is **extremely dangerous**. *– Indirect NegE, Target_Council: Country, Ukraine, Target_Intermediate: Other*

(UNSC_2014_SPV.7154, Russian Federation)

From the first sentence, a potential Target_Council can be inferred since *Turchynov* is a representative of Ukraine. The second and third sentence contain a negative evaluation. But is the critique mainly targeted at situation or the representative of Ukraine? For the first NegE, we maintained that we cannot be entirely sure that Turchynov is held responsible for the negatively evaluated situation, hence it is indirect. The same holds for the second NegE. In the next example, we decided for Indirect NegE since one sentence prior to it, *pro-Russian militants and separatists* were mentioned.

(9) At the same time, when the **existence of a State is put in danger**, we understand and support the right of Ukraine to defend itself **in the face of external aggression** and to tackle **militant separatism and continuous provocations** in order to protect the State and its population from a **further escalation of violence**. – *Indirect NegE, Target_Council: Underspecified, Target_Intermediate: Non-Governmental Group* (UNSC_2014_SPV.7154_spch005, Lithuania)

The question here is whether "external aggression" can be interpreted as meaning the Russian Federation. We decided that it is not, since "pro-Russian" does not automatically mean that the actions described are supported by the Russian Government.

6.3.3. Lexical Markers for NegE

In the next example, our annotators disagreed on whether "two different standards" is a NegE:

(10) We also call on Russia to fully assume its role as a permanent member of the Security Council. – *no Conflict*

Russia is a guarantor of peace and security in the world, both in the larger world and in its immediate neighbourhood. – *no Conflict*

There can not be **two different standards**. – *Direct NegE, Target_Council: Country, RF*

(UNSC_2014_SPV.7154, France)

Grounding annotations in lexical markers has generally proven very helpful in the annotation procedure. However, since it is not feasible to provide a complete list of lexical markers that qualify as negative evaluation, the decision was often difficult.

We think that the examples given above demonstrate the challenge of annotating diplomatic language, where the discrepancy between what is said and what is meant can be rather large, and purposefully so.

7. Conclusion and Future Work

We presented a new annotation scheme for conflicts in the UNSC, which can be used to analyze information about countries' positions and alliances, and the degree of direct confrontation in the speeches. The framework is closely connected to sentiment analysis (and in principle also to disagreement detection), but is adapted to the particularities of pre-written diplomatic speeches. Although the task is far from straightforward, we gain a moderate to substantial inter-annotator agreement.

In the classification task, the lexicon-based sentiment classifier used mainly by the Political Science community (Lexicoder) for the binary classification task conflict/no conflict scores 0,71 f_1-weighted, which is only slightly less than the results with our tested BERT-based classifiers `distilbert-base-uncased`, `sentiment-roberta-large-english` and `roberta-argument`.

By defining linguistic markers and precisely formulated guidelines for conflicts for our annotators, we were able to detect a sizable amount of conflict statements, also for the WPS agenda with less obvious disputes than in the Ukraine debates.

In the next step, experts on international diplomacy will examine our annotations to estimate a degree of conflict, but also to point at conflicts we were not able to detect.

Having an annotated dataset of conflicts opens up various possibilities for follow-up projects, both quantitatively and qualitatively. For classification models, we would like to expand the annotations for having more training material. Looking at more agenda items with possibly different types of conflicts could help the models to generalize and perform better. Using EDUs as granularity for our annotations, we laid the groundwork for subsequent analysis that looks at discourse structures of the debates. We have started to annotate our data following the RST framework (Mann and Thompson, 1988), and our next step is to look at rhetorical strategies used by diplomats to express conflicts in more detail.

8. Limitations

The annotation framework is language-agnostic, but for the experiments we only used it for speeches given in English or translated into English. Translation at the UN is highly institutionalized and has established a set of translation norms to ensure a high quality of translations (monitoring programs, terminology, proof reading).[14] Furthermore, when working at important events like Security Council meetings, interpreters are often allowed to prepare their translations and have access to information about the proceedings prior to translating the speeches, allowing them to familiarize themselves with the concepts and terminology of the debate.[15] Despite the high quality of the translations, they can generate problems of cross-cultural misunderstandings (Cohen, 1991). Translations can therefore change the character of the statements and cause potential conflicts to be lost or exaggerated.

[14] https://jostrans.org/issue09/art_cao.pdf

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

Since the annotation procedure is relatively expensive, we currently cannot analyse the nature of speeches (and the amount of conflicts in them) from the same country over a longer period of time in great detail. We hope that by annotating more debates, such longitudinal analysis can be performed on the future.

Furthermore, choosing debates from two agenda items from a limited time frame might infuse bias to the annotations. As already stated in Section 7, we plan to expand the annotations to diversify the annotated data and conduct further experiments. The annotation guidelines were discussed on different occasions with an International Relations expert's community, but the annotations themselves were conducted by students from our computational linguistics department, who had a background in linguistics. The annotators were familiar with the texts from another project, and our guidelines are linguistically motivated. Nevertheless, we will continue discussing the dataset with experts to detect conflicts that our annotators were unable to detect and discuss potentially ambiguous cases.

9. Acknowledgements

The paper was supported by the Deutsche Forschungsgemeinschaft (DFG), project (448421482) "Trajectories of Conflict: The Dynamics of Argumentation in the UN Security Council“. We thank our annotators Dietmar Benndorf and Costanza Rasi for their work and their valuable feedback on the annotation guidelines.

10. Bibliographical References

Kelsey Allen, Giuseppe Carenini, and Raymond Ng. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1169–1180, Doha, Qatar. Association for Computational Linguistics.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Conference on Empirical Methods in Natural Language Processing*.

Raymond Cohen. 1991. *Negotiating Across Cultures: Communication Obstacles in International Diplomacy*. Number 31 in Negotiating Across Cultures: Communication Obstacles in International Diplomacy. United States Institute of Peace.

Susan Conrad and Douglas Biber. 2000. *Adverbial Marking of Stance in Speech and Writing*, pages 56–73. Oxford University Press, UK.

Christine De Kock and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emanuel Deutschmann, Jan Lorenz, and Luis G. Nardin. 2020. Advancing conflict research through computational approaches. In Emanuel Deutschmann, Jan Lorenz, Luis G. Nardin, Davide Natalini, and Adalbert F. X. Wilhelm, editors, *Computational Conflict Research*, Computational Social Sciences, pages 1–19. Springer International Publishing.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 669–676, Barcelona, Spain.

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.

Kristian Skrede Gleditsch. 2020. Advances in data on conflict and dissent. In Emanuel Deutschmann, Jan Lorenz, Luis G. Nardin, Davide Natalini, and Adalbert F. X. Wilhelm, editors, *Computational Conflict Research*, Computational Social Sciences, pages 23–41. Springer International Publishing.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Maia Kutateladze. 2020. Criticism in diplomatic letters. *Journal in Humanities*, 9(2):71–75.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Seraphine F. Maerz and Cornelius Puschmann. 2020. Text as data for conflict research: A literature survey. In Emanuel Deutschmann, Jan Lorenz, Luis G. Nardin, Davide Natalini, and Adalbert F. X. Wilhelm, editors, *Computational Conflict Research*, Computational Social Sciences, pages 43–65. Springer International Publishing.

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. Martin and P.R.R. White. 2007. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan UK.

Deepak Nair. 2019. Saving face in diplomacy: A political sociology of face-to-face interactions in the Association of Southeast Asian Nations. *European Journal of International Relations*, 25(3):672–697.

Masanori Oya. 2011. Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*, pages 313–316, Hong Kong.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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 Schönfeld, Steffen Eckhard, Ronny Patz, and Hilde van Meegdenburg. 2019. The UN security council debates 1995-2017. *CoRR*, abs/1906.10969.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.

Giuseppina Scotto di Carlo. 2012. The language of the UN: Vagueness in security council resolutions relating to the second gulf war. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 26.

Hannah Slavic and Jovan Kurbalija. 2001. Language and diplomacy: Preface - diplo resource.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 26–34, Antwerp, Belgium. Association for Computational Linguistics.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. Annotation guidelines for rhetorical structure.

Mirco Schönfeld Steffen Eckhard, Ronny Patz 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.

Maite Taboada. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347.

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

Justine Zhang, Cristian Danescu-Niculescu-Mizil, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Annual Meeting of the Association for Computational Linguistics*.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.

11. Language Resource References

Prasad, Rashmi and Webber, Bonnie and Lee, Alan and Joshi, Aravind. 2019. *Penn Discourse Treebank Version 3.0*. ISLRN 977-491-842-427-0.