Amendment RST Annotation Guidelines for Diplomatic Speeches in the UNSCon: EDU Segmentation and RST Labeling

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

Abstract

Amendment for RST annotations on the UNSCon Corpus.

Contents

1   Introduction

The UN Security Council Conflict Corpus (UNSCon) consists of transcripts of speeches given in the UN Security Council (UNSC) and includes multiple layers of annotations, where the goal is to examine how diplomats discuss and argument during meeting. One of the annotation layers is a task we call Conflict detection, from which the corpus' name is derived from. It includes a content-based analysis of verbalized disputes and negative evaluation of meeting participants. The second layer of the corpus is the analysis of discourse structures of speeches looking at *how* things are expressed. We base this analysis using Rhetorical Structure Theory (RST, Mann & Thompson (1988)). The annotation of rhetorical structures for RST includes several steps, from which discourse segmentation and discourse tree-building are the two primary components. Both steps were already defined by guidelines for existing corpora including RST annotations, like the RST-Discourse Treebank (RST-DT, (Carlson et al., 2002)), the RST layer in the GUM corpus (Zeldes, 2017), and the English RST guidelines for the Potsdam Commentary Corpus (PCC-RST, (Stede et al., 2017)). The segmentation of text spans into Elementary Discourse Units (EDUs) in UNSCon are based on RST-DT. Based on the particularities of the corpus, we included additional rules, to provide consistent segmentation 2.

Another steps in RST is choosing rhetorical relations between EDUs (which defines which function a EDU holds in relation to another EDU) and weighting the EDUs based on their importance, which creates a a tree-like structure of the document examined. For choosing RST labels we mainly follow the PCC-RST guidelines as defined by Stede et al. (2017), but modify it slightly in expanding the label set. Another goal was to to provide interchangeability between the different variants of RST,

namely RST-DT and GUM. Consequently, section includes a mapping of the labels between our label set and RST-DT and GUM respectively.

We include a variety of examples from the corpus to put an emphasis on the special type of text we are dealing with, which are diplomatic speeches with a tendency to long and complex sentence. Inside the examples, the parts we want to highlight are in bold. Orthographic mistakes coming from the original data are marked with *sic*.

2   EDU Segmentation

The first step in characterizing the discourse structure of a text is to determine the Elementary Discourse Units (EDUs), which are sentences or smaller. Sentences are always used as separate segments for RST analysis, but are sometimes still to coarse-grained for rhetorical units, that is why sentences can be segmented to smaller units which are often subordinate clauses. The EDU segmentation rules for UNSCon basically follow the ones defines for the RST-DT and GUM corpus.[1] Nevertheless, based on that we work with diplomatic texts and transcripts of speeches which were often translated to English, and which sometimes include mistakes coming from parsing errors (the speeches are parsed transcripts of PDF files) we felt that some rules already formulated in RST-DT should be further explained and expanded, which we will show on examples from the corpus. For example, we could not solely rely on comma to extinguish requisite and unrequited segmented relative clauses (as explained in RST-DT) since they could be missing from an transcription error. We also point to the interpretation space the segmentation guidelines left, meaning that even though the RST-DT segmentation guidelines include precise segmentation rules mainly based on syntactic rules, content interpretation also play a part in the segmentation (especially for relative clauses). Again, we will provide some examples for those cases.

2.1   General Notes on Segmentation

2.1.1   Segmentation Based on Punctuation

Since EDUs are defined as sentences or smaller, we segment them by the sentence's final punctuation. In general, punctuation such as commas, semicolons, colons and periods are part of the preceding EDU. An exception to this rule are **paratheticals** and **dashes** to mark for example insertions:

> [The violation of international law is so obvious at this point] [that one almost feels pity at seeing Russian diplomacy] [**- so formalistic, so finicky in its respect for proprieties and its invocation of texts -**] [struggling to find a legal basis for the coup.] (UNSC_2014_SPV.7138_spch005_sentsplit_France.txt)

2.1.2   Segmentation Based on Automatic Sentence Splitting

Generally, we formulate the rule for the UNSCon as an extension to RST-DT:

> Rule 1: An EDU is smaller or as long as one sentence. For cases when automatic segmentation lead to false sentence segmentation and no given rhetorical relation can be chosen, we use the relation SAME-UNIT.

The sentence splitting for the UNSCon was conducted automatically using SpaCy sentence splitting module. The sentence-splitted corpus is our reference corpus for any further annotation layer in terms of character offsets. In order to keep the corpus generation reproducible and keep the character offsets consistent, we keep the generated segmentation even if mistakes occur. Sentence segmentation mistakes happen rarely and especially around parentheses, where we would include an EDU segmentation in most of the cases anyway (see section 2.2.9). Instead of manual correcting, we

---

[1]The GUM corpus basically follow the segmentation rules from RST-DT providing a comprehensive overview of their guidelines on https://wiki.gucorpling.org/gum/rst.

align the EDU segmentation and SAME-UNIT as relation if there is no other existing relation that can be established between the units.

2.1.3 Dealing with Incorrectly Parsed Documents

The corpus we are working with is based on the published UNSC dataset by Schönfeld et al. (2019) which parsed the original PDF files provided by the UN Digital Library into text files. The process sometimes lead to mistakes in the output leading to false or missing characters or spaces like in the next example:

> [Out of all ofthis (sic)] [- their confused flurry of activity,] [the speeches that are denied] [as soon as they are given] [and the arguments that are forgotten] [as soon as they are made i (sic)] [only force remains.] (UNSC_2014_SPV.7138_spch005_France_03.txt)

We handled those mistakes in stating that if it is possible to interpret an appearing orthographic mistake, we segment the text as if it were corrected. Therefore we interpret "i" in the example as dash and segment the inserted clause accordingly.

As a general preprocessing step for the UNSCon, we excluded everything (like the name of the speaker, the language spoken) that is not part of the speech itself, so that for the majority of files only the speech's content is processed. Generally one file should only include one speech. Nevertheless, there are cases where mistakes in the OCR processing lead to additional content coming from previous or subsequent speeches, like in the next example. For those cases we include the rule:

Rule X: For documents in the corpus including additional text fragments from other speeches, we segment the text that does not belong to the speech and connect it with the main text using the TOPIC-COMMENT relation we borrow from GUM and RST-DT.

In the next example, the introduction of the president (the first EDU) and the Information about the next speaker (second and third EDU) are grouped and via TOPIC-COMMENT connected to the rest of the speech.

> [[[I now give the floor to the representative of Brazil.] $->$ TOPIC-COMMENT
>
> [[Mr. De Aguiar Patriota]$<-$ BACKGROUND [(Brazil)]]]] $->$ TOPIC-COMMENT
>
> [Let me congratulate Angola on having organized this debate under the guidance of its Minister for Family and the Promotion of Women, Ms. Maria Filomena Delgado.] (UNSC_2016_SPV.7658_spch029_sentsplit_Angola.txt)

In the next example, at the end of the document, the text of the president introducing the next speaker is falsely included. We therefore segment those parts and connect them using TOPIC-COMMENT as relation.

> [[We will continue to advance these inseparable agendas at all levels of the United Nations.] $<-$ TOPIC-COMMENT [[The President:] $->$ TOPIC-COMMENT [I give the floor to the representative of India.]]] (UNSC_2016_SPV.7658_spch029_sentsplit_Angola.txt)

2.2 Segmentation Based on Syntactic Rules

As stated, making use of syntactic rules for the segmentation gives unifying rules for segmentation, whereas a segmentation only based on semantics would lead to (even bigger) interpretation scope. The grammatical construction of subordination indicates that one part of the sentence is secondary (or subordinate) to another part. Since in RST we also weight the importance of discourse units into more important (nucleus) and less important (satellite) ones, we can make use of rules based on

detecting subordinate or main clause. We will show different forms of subordinate clauses and their segmentation rules in: section 2.2.2 for relative clauses, section 2.2.5 for cleft constructions, section for Same-Unit constructions, and section 2.2.6 for complement clauses.

In English, coordination is a way of connecting ideas that are roughly equal in importance, which are usually segmented as EDUs. For coordinated clauses it is straightforward and they are generally segmented. Nevertheless, coordination can lead to elliptic construction, where for example the main verb or the subject is shared among the clauses, for which we look at the segmentation rules in section 2.2.7 for coordination.

2.2.1 Clauses

> Rule X: We segment finite clauses as EDUs, for non-finite clauses we segment subordinate clauses in elliptic constrictions.

As we have already noted, sentences are the basic unit of discourse, but they are are often segmented into smaller units, where *clauses* are often the main unit segmented.[2]

In grammar, we can distinguish between finite clauses that contain a subject and a finite form of a verb, and non-finite clauses containing non-finite form of a verb as its center element. Finite verb forms in English grammar show tense, person and number. Non-finite verb forms do not show those attributes, typically they are:

- infinite forms with and without to (*to go*, *go*)
- *-ing* forms (gerunds like *going*)
- *-en* forms (participles, stolen)

Usually non-finite verbs are found in elliptic subordinate clauses and if the subject is interfered from the main clause like in the prepositional phrase: "I had something to eat *before leaving*". We also segment those clauses (see RST-DT sec. 2.9.2, p. 21). Non-finite clauses are often found after subordinating conjunctions like *after, although, though* and *if*, which often points to a rhetoric relation, therefore they are mostly segmented.

2.2.2 Relative Clauses

A common form of subordination appearing often in the speeches of the UNSCon is a relative clause, which is a clause that usually modifies a noun phrase and is introduced by a *relative pronoun* (which, that, who, whom, whose), *a relative adverb* (where, when, why), or a *zero relative* (leaving out the relative pronoun). It is a *postmodifier*, meaning that it appears after the noun phrase it modifies, but it doesn't always need to immediately follow the noun phrase that it modifies.

2.2.2.1 Relative Clauses Handled in Other RST Frameworks

In English grammar, relative clauses are usually divided into **restrictive and non-restrictive relative clauses**, while this distinction is important to understand the segmentation rules as defined in PCC-RST. A nonrestrictive relative clause is a relative clause that adds nonessential information to a sentence. It does not limit or restrict the noun or noun phrase it modifies and is usually set off from the rest of the clause by commas, dashes or similar typography. In contrast, restrictive relative clauses limits the noun that is referred to and is necessary for the meaning of the sentence.

For example, in the following constructed sentence

The paint *which Mary bought at the hardware store* was bright red.

---

[2]from PCC-RST guidelines: "[...] discourse segments are clauses and sentence", p. 4, sec. 3.1)

the restrictive relative clause limits which paint we're referring to, namely to paint which Mary bought at the hardware store.

In contrast, in the sentence:

> The paint, *which Mary bought at the hardware store,* was bright red.

the nonrestrictive relative clause, does not restrict the reference of the noun paint; it is not information that distinguishes the paint from other paint. That Mary bought this paint at the hardware store is simply incidental information. In PCC-RST, restrictive relative clauses do not necessarily state an discourse unit. Nonrestrictice relative clauses referring to a *discourse entity* and not only restricting the noun, and if they are set off from the rest of the clause by typography, are on the other hand segmented.

In contrast, GUM does not make this distinction and states that generally relative and reduced relative clauses that modify a noun are segmented, including the example "[We met the woman] [who lived in the house]" where only the noun *woman* is modified. RST-DT also generally segments relative clauses (sec. 2.9.1., p.21) if they contain a verbal element. They are not segmented if they include only an adjective (-able) without a verbal element.

2.2.2.2   Relative Clauses Handled in UNSCon

For segmenting relative clauses in the UNSCon, in contrast to PCC-RST but following GUM and RST-DT, we implemented the following general rule:

Rule X: *We generally segment relative clauses and do not distinguish between restrictive and nonrestrictive ones. Also, we do not rely on commas, parentheses or other typographical features marking a separation.*

We decided against relying on typographical features that mark a separation because we are dealing with translations and transcripts of speeches, which – in our experience – are prone to orthographic errors, also due to incorrect PDF parsing of the original PDFs. This rule was also implemented due to our experience that especially when we faced complex and nested sentence constructions with several embedded relative clauses which were refered to again, we felt that only segmenting nonrestrictive relative clauses would be to narrow and we would loose several discourse units. The next example includes a segmented relative clause *who refused to subscribe to their sloganeering* which is segmented:

> [I would recall] [that the collaborators and the followers of Bandera killed not only Jews, Poles and Soviet soldiers] [but also Ukrainians] [who refused to subscribe to their sloganeering.] (UNSC_2014_SPV.7165_spch015_sentsplit_Russian_Federation.txt)

An example for a segmented zero-relative restrictive relative clause (leaving out the relative pronoun *which*) is in the next sentence, expressing an ELABORATION relation.

> [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.] (UNSC_2014_SPV.7138_spch002_sentsplit_Russian_Federation.txt)

Nevertheless, we included some rules restricting the segmentation to some cases:

Rule X: *Following RST-DT, we state that clauses do not segment reduced relative clauses that contain an adjective (-able) without a verbal element. There are content-based exceptions, see section 2.3.*

Sometimes is tricky to distinguish between relative clauses and a complement of noun (see section 2.2.6). In the next example, the propositional phrase starting with *to all* contains a noun complement

defining in more detail "all", and therefore the complement starting with *who have the responsibility* is therefore not segmented:

> [With this resolution, the Security Council is sending a strong message] [to all who have the responsibility to eliminate this scourge,] [and that responsibility is collective.] (UNSC_2016_SPV.7643_spch014_sentsplit_Spain.txt)

In the next example, the enumerations are segmented (as stated in RST-DT rule 2.11.5: "Text fragments followed by colons are treated as separate EDUs, even when the fragment is a word or phrase, as long as the text that follows the colon provides further elaboration on the topic introduced by the colon:[...]"), but the relative clauses are not segmented since the segments referred to (in italics) do not contain a verb. Also the next example shows the complexity one sentence possible in the corpus:

> [Nothing is lacking from *those* things that we hoped were relegated to the vaults of history:] [the military maneuvres at the borders,] [*the henchman* who yesterday was no one and today has suddenly been designated President of the Crimea,] [*the occupation* denied against all evidence,] [the torrents of propaganda,] [and now *the charade of a referendum* that is not only illegal, rushed and without electoral campaigns] [but is reduced to a choice between two ways of saying yes.]
>
> UNSC_2024_SPV.7138_spch005_sentsplit_France.txt

The first EDU in the example is not segmented because of the demonstrative pronoun *those* is establishing a close connection between the subordinate clause "Nothing is lacking from *those* things" and the ordinate clause "that we hoped were relegated to the vaults of history:". At this point we introduce another new rule for the UNSCon segmentation guidelines:

> Rule X: If there is a close (and necessary) connection stated between the noun and the phrase which specifies it, we do not segment it. This can be stated through using a **demonstrative pronoun** (this, that, these, and those) or using *any* in the main clause which is referring to a noun phrase in the relative clause.

> [Secondly,] [in the meanwhile,] [all parties should refrain from taking *any action* that may further escalate the situation.] (UNSC_2014_SPV.7138_spch009_sentsplit_China.txt)

The next example includes relative clauses not segmented since there is a relative pronoun. Only the prepositional phrase (*for arming those who use those weapons*) modifying a noun and including a main verb is segmented (see GUM, sec. "Relative and modifier clauses").

> [Those who are responsible] [for arming those who use those weapons] [bear the responsibility for this conflagration.] (UNSC_2014_SPV.7219_spch011_sentsplit_France.txt)
>
> [It is for that reason that the United States has supported mobile courts in the Democratic Republic of the Congo,] [which has led to 1,924 trials and 1,336 convictions of gender-based violence since 2009.] (UNSC_2016_SPV.7658_spch013_sentsplit_United_States_Of_America.txt)

In the second example, the demonstrative pronoun *that* is a marker for not segmenting. Implicitly this rule it is included in the RST-DT guidelines, which states that if for the meaning the subordinate clause is necessary, we do not segment it. This is the case for clauses with demonstrative pronoun referring to a noun in the relative clause. Sometimes this close connection can also be stated using a definite article "the" which could be replaced with a demonstrative pronoun. This is the case in *the principles* in the next example:

> [The delegation of Argentina reiterates] [that it is essential that we adhere to *the* principles that we subscribe to as Members of the United Nations.] (UNSC_2014_SPV.7154_spch013_sentsplit_Argentina.txt)

Another segmentation rule for relative clauses we include is:

> Rule X: *The clause the relative clause is referring to needs a main verb, an auxiliary verb is not enough (be, have, do, can / could, may / might, will / would, shall / should, must, needn't, mustn't, ought to). An exception are correlatives consisting of two discourse markers both in the subordinate and main clause.*

Therefore, next example is not segmented:

> [It is a secret to no one that the Russian Federation will vote against the draft resolution.] (UNSC_2014_SPV.7138_spch002_sentsplit_Russian_Federation.txt)

As stated, this rule does not apply to all subordinate clauses, which are correlatives that consists of a combination of two discourse markers, in the subordinate clause and the other in the ordinate clause (see RST-DT, sec. 2.8, p. 18). Many of those correlatives have a comparative or contrastive function. Constructions of this type should be broken into separate EDUs if both include a verbal element, where the subordinate clause can contain either a main verb or an auxiliary:

(110) [It was as easy] [as collecting shells at Malibu.] (115) [Moreover, "it's a lot **cheaper** and **quicker** to buy a plan] [**than** to build one."]

Free relative clauses are relative clauses that contain the reference to the noun phrase in itself and are lacking a noun as head.

> Rule X: Following GUM (see sec. "Free relatives"), free relative clauses are not segmented.

An example is: [I found what you were looking for.]

Usually relative clauses begin with a relative pronoun like *who*, *which*, and *that*, referring to the subject. In the UNSCon, there are cases, where a dash can be interpreted as left out for a relative pronoun (as "which are out and out violations of that agreement"). This is a hint for the mixed registers used during the speeches, from predefined sentences to free-speeches with left outs, which can complicate the EDU segmentation.

> [However, in just a matter of days, there was a resumption ofthe so-called counter-terrorist,] [but actually punitive, operations] *[- out and out violations of that agreement.]* (UNSC_2014_SPV.7165_spch015_sentsplit_Russian_Federation.txt)

Another rule for relative clauses is:

> Rule X: *We follow RST-DT in stating that if a relative clause starts with a quantifier, we include that as part of the relative clause.*

2.2.3 Other Modifying Clauses and To-Infinitives

Next to relative clauses, there are other modifying elements that can be segmented as EDUs.

> Rule X: We segment *to-infinitives modifying a noun*, usually using ELABORATION, or sometimes PURPOSE relation (taken from GUM).
> *That*-clauses modifying a noun and prepositional clauses are also segmented. For *prepositional clauses*, the headword of the adnominal clause must be a verb (typically a gerund in -ing). We do not segment modifying clauses referring to a verbal element or adjective (and therefore not to a noun phrase).

The next invented sentences serve to exemplify the differences for to-infinitives:

[They made the decision] [to go.] (explanation: "to go" modifies the noun)

[We had a chance] [to succeed]. (explanation: to succeed modifies the noun "chance")

[We <u>have tried</u> our best <u>to ensure</u> the people will succeed.] (not segmented since it is a verb complement)

Following GUM and PCC-RST, object clauses (complements, see section 2.2.6) are not segmented, therefore *to ensure* which may take a clause as object is not an marker for a EDU segmentation.

We also do not segment attribution (or reported clauses) from to-infinitives (see section 2.2.6.2).

2.2.4 Same-Unit Constructions

> Rule X: Following RST-DT and GUM we treat relative clauses, nominal post-modifiers, appositives, parentheticals as embedded EDUs which can interrupt one EDU. When they break up an discourse unit we connect the interrupted units with SAME-UNIT relation. For interrupting constructions without a verb phrase, we additionally state that they must be typographically marked. We also state that adjectives and adverbs standing alone but being marked by commata, cannot break up an EDU.

We differ here from PCC-RST, who do not include the relation label SAME-UNIT, since Same-Unit constructions occurs relatively often in the speeches. An example for a relative clause which is embedded into the EDU is:

[But, again, a major debate] [that includes criticism] [is a sign of democracy.] (UNSC_2014_SPV.7165_spch016_sentsplit_Lithuania.txt)

While in the examples in RST-DT there is always only one embedded unit breaking up an EDU, we see this quite often happen in the speeches in the UNSCon, that there are complex constructions breaking up an EDU. In the example below, the relative clause is embedded in the clause *Undertakings to allow access must be upheld*, where the noun of the clause is again specified by a relative clause. We segment both and generally include the rule:

> Rule X: We allow hierarchically structured embedded discourse units, like in the following example, where the embedded unit is segmented in three:

[Undertakings] [**brokered by the trilateral contact group with armed separatists**] [**who control the territory in which the crash occurred**] [**to allow access**] [must be upheld.] (UNSC_2014_SPV.7219_spch009_sentsplit_Australia.txt)

RST-DT is relatively free in defining possible units that can break up an EDU, as noun postmodifiers or participal constructions, without restraining it having a verb as head (section 2.9.7, p.23). To keep the rule with RST-DT consistent, but to avoid unclear cases, we include as rule that constructions without a verbal element can be segmented, if they are typographically marked. This partly contradicts our rule stated for relative clauses that we don't want to rely too much on typography for the UNSCon (see section 2.2.2.2), but here we wanted to implement countermeasures for excessive EDU segmentation (since they are used a lot in the speeches) and felt that it was not severe if we didn't catch all of these insertions.

We also do not segment adjectives or adverb standing alone, which would not be break up if they would be at another place of the sentence (In the next example, the adverb would normally be in front of the verb "saying" and not segmented).

[In Washington, DC, they are saying, *uncritically,*] [that the buildings have been legally rented.] (UNSC_2014_SPV.7165_spch015_sentsplit_Russian_Federation.txt)

We also do not with column segmented names for the same reason (the next example could be reformulated in *Donetsk People's Governor Pavel Gubarev remains in jail.*).

> [However,] [instead of that,] [the People's Governor of Donetsk, *Pavel Gubarev,*] [who never had weapons in his hand,] [remains in prison.] (UNSC_2014_SPV.7165_spch015_sentsplit_Russian_Federation.txt)

2.2.5 Cleft, Extraposition

Rule X: *Like others, we do not segment cleft nor extrapositions.*

A cleft is a complex sentence (having a main and a subordinate clause) that has a meaning that could be expressed by a simple sentence. It is a a construction in which some element in a sentence is moved from its normal position into a separate clause to give it greater emphasis. For clefts, one can separate between "it-clefts" (*It was Mia who read the book.*) and "wh-clefts" (*What Mia needed was a book.*, also often called pseudo-cleft). The next example could be expressed with a simpler sentence (The West must determine (...)) but in order to emphasize the subject it was taken to the beginning:

> [*It is the West* that must determine the opportunity for avoiding civil war in Ukraine.] (UNSC_2014_SPV.7154_spch004_sentsplit_Russian_Federation.txt)

Extrapositions are similar to cleft (since both can have a meaningless dummy pronoun, *it*, as Subject), but instead of breaking a sentence into two parts, it just moves the clause to the front position. The sentence *That the draft was immediately criticized (…) was no surprise* is reversed from the next example:

> [*It is no surprise* that the draft was immediately criticized in the Rada by all sides.] (UNSC_2014_SPV.7165_spch015_sentsplit_Russian_Federation.txt)

A second example is the first EDU in the next sentence, which is not segmented since e the subject (in italics) has been extraposed to the that-clause:

> *[It is also understandable that the enjoyment of the right to self-determination]* [involving separation from an existing State] [is an extraordinary measure,] [applied] [when further coexistence within a single State becomes impossible.]

2.2.6 Complement Clause

PCC-RST states that complement clauses are no EDUs. A complement clause is a subordinate clause that serves to complete the meaning of a noun or verb in a sentence. Taken from generative grammar theory (Hurford, 1994), we can refer to various closely related kinds of subordinate clause:

1) Subordinate clauses that serve as **direct (and therefore necessary) object of a verb**, such as believe, say, tell, know, understands. There are cases when they are segmented as ATTRIBUTES, see section 2.2.6.2. (I can assure you *that Bob is quite proficient at electronic data manipulation.*)

2) Subordinate clauses which **modify nouns** (such as story, rumour, and fact) and **adjectives** (such as proud, happy, and sad). **We do not segment those complements.** See section 2.2.6.3.

3) Subordinate clauses which **act as the subject of sentences** with predicates as *be a pity, be a nuisance, be unfortunate, seem,* and *happen.* **We do not segment those complements.**

There are two exceptions we adopt from RST-DT, when complements can be segmented: for complements of Attribution, and if complements start with a strong discourse marker. We will start with defining those strong discourse markers taken from RST-DT and expand the list.

2.2.6.1 Strong Discourse Markers

> Rule X: Complements are segmented if there is a strong discourse marker RST-DT defines a finite set of strong discourse cues which is:
>
> - *because*; **we additionally include:** *in order to*, *to*
> - *as a result of*
> - *in spite of, despite*
> - *regardless* (but not *with regard to*)
>
> - *irrespective*
> - *without*
> - *according to*
> - *not only … but also*

The first EDU in the sentence does not include a verb phrase but is segmented since it is separated with the strong discourse marker "despite" using CONCESSION relation.

> [*Despite* the unilateral approach to the negotiations on the draft resolution,] [we have chosen not to vote against it.] (UNSC_2016_SPV.7643_spch005_sentsplit_Egypt.txt)

Another example not listed in the strong discourse markers is nevertheless segmented, since we saw *not just* as a strong cue for a contrastive rhetorical relation (here CONCESSION). We had more cases, where we segmented a phrase without a verb phrase because we saw a strong cue for an contrastive relation.

> [When asked "What was your role?",] [she will be able to say] [that she was at the forefront of the United Nations action] [to prevent conflict and resolve crisis and represent the entire world,] *[notjust(sic) half of it.]* (UNSC_2016_SPV.7658_spch012_sentsplit_United_Kingdom_Of_Great_Britain_And_Northern_Ireland.txt)

2.2.6.2 Complements of Attribution

ATTRIBUTION, sometimes also called a *reporting clause*, serves for the identification of the speaker or source of something reported. For RST annotations, it is uninteresting if the attribution refers to direct or indirect speech. It consists of a subject and a verb of speaking or writing, and optionally other related information as adjectives (how something is said).

PCC-RST does not segment Attribution since the reported unit is syntactically a noun clause complement of the reporting verb (see PCC-RST, p. 6, sec. 3.2.3). Also they state that they doubt, that they enter into discourse relations of the type that RST attempts to capture.

Nevertheless, for the UNSCon, we are interested in analyzing how often speakers refer to other sources or themselves in their line of argumentation. Therefore, we decided to include ATTRIBUTION into our relation set and stay therefore closer to the guidelines as defined in RST-DT, which has implications for segmentation, since otherwise these clauses would be not segmented.

Rule X: Attributions are only marked if the following conditions are met (see RST-DT, 2.4, p.7):

1) There is an explicit source of the attribution (no passive construction); and

2) The attribution predicate takes a clausal complement that is *not infinitival.*

Since this is not explicitly stated in RST-DT, we include the rule that:

3) Clausal complements having a verb taking a gerund-form (-ing) (and therefore having the function of a noun), are not segmented.

This means that the content of the reported message (in the separate clause) should include an inflected verb. Therefore, if the complement includes a to-infinitival, we do not segment, like in the next example where *(re)affirmed* is once not segmented and once not.

[And yet,] [in defiance to the Budapest Memorandum,] [whereby **the Russian Federation reaffirmed its obligation to refrain** from the threat or use of force against the territorial integrity or political independence of Ukraine,] [contrary to the Treaty on Friendship, Cooperation, and Partnership between Ukraine and the Russian Federation of 1997,] [**which affirms**] [**that both countries agree to respect each other's territorial integrity**] [and confirm the viability ofthe *(sic)* borders existing between them,] [and in violation of the obligations and commitments undertaken by Russia] [during its accession to key regional and international organizations,] [Ukraine is about to be carved up.] (UNSC_2014_SPV.7138_spch007_sentsplit_Lithuania.txt)

The next example, the first EDU is not segmented since *duelling* in the object complement is in this case an adjective:

[United Nations monitors have also reported duelling rallies in Kharkiv today,] [resulting in injuries to some 50 persons.] (UNSC_2014_SPV.7154_spch002_sentsplit_UN.txt)

The next example the complement after "disclaimed" includes a gerund "trying" and a to-infinitive "to annex", and is therefore not segmented.

[Russian leadership has disclaimed any intention of trying to annex Crimea, (...)] (UNSC_2014_SPV.7138_spch004_sentsplit_United_States_Of_America.txt)

The question of what is an ATTRIBUTION is non-trivial (as already noted in Stede et al. (2017)).

Many of the speeches include formal formulations that are serving for alignment and expressing gratitude than for reporting purposes. We do not see these constructions of greetings and gratitude as attributions.

[I thank Under-Secretary-General Feltman for his briefing and his warning to the Council.] (UNSC_2014_SPV.7165_spch014_sentsplit_Australia.txt) – not segmented since briefing and warning are nouns. [Australia strongly supported calls] [for the holding of tonight's briefing] [to give the Council an up-to-date sense of the continuing destabilization and increased tensions in eastern Ukraine,][...] (UNSC_2014_SPV.7165_spch014_sentsplit_Australia.txt) – not segmented since holding is a noun

[China expresses its appreciation to you, Minister Delgado, for presiding over this meeting.] (UNSC_2016_SPV.7658_spch015_sentsplit_China.txt) – not segmented since prepositional phrase is an clausal objects, which are not segmented (RST-DT, p. 5, sec. 2.2)

In the next example the first sentence could be paraphrased with a formulation which would lead to a segmented Attribute like: "The vote today says that Russia...".

[**The resounding message from today's vote is** that Russia stands isolated in the Council and in the international community.] (UNSC_2014_SPV.7138_spch006_sentsplit _United_Kingdom_Of_Great_Britain_And_Northern_Ireland.txt)

Nevertheless we did not segment it since the RST-DT guideline explicitly state that clausal complements of *attribution verbs* (see RST-DT, 2.4, p.7) are segmented.

We came to the same conclusion for the next two examples, where in the first one the complement *It is our believe* is not segmented. Also, we interpret the construction *that resolution 2272 (2016) will send a clear message (...)* as the noun complement for the noun *belief*:

[It is our belief that resolution 2272 (2016) will send a clear message (...).] (UNSC_2016_SPV.7643_spch017_sentsplit_Japan.txt)

[Pursuant to those declarations,] [real Ukrainians should have collaborated with Hitler] [to establish a new order in Europe.] (UNSC_2014_SPV.7165_spch015_sentsplit_Russian_Federation.txt)

The first EDU in the second example is segmented, not using ATTRIBUTION but a causal relation CAUSE.

Next to verbs expressing speech acts, RST-DT also include verbs expressing cognitive processes as "to think". In our experience, this opens up to several questions when annotating, if the complement is really an attribution. Although RST-DT gives example verbs for both cases of ATTRIBUTION the list cannot be comprehensive and there are still verbs left where we have a scope of interpretation if they are part cognitive or speech acts. We decided to expand the list of attributive verbs based on what we often found in the UNSCon. The following lists includes verbs marking speech and cognitive acts as listed in RST-DT, but expanded with candidates found in our corpus.

We included the following verbs as candidates for **speech acts**:

- according to
- advise
- announce
- ask
- assert
- (dis)claim
- confirm
- condemn
- declare
- demand
- deny

- explain
- indicate
- make it clear
- mention
- note (in a sense of *saying*, *making a remark*, but not in sense of *noticing*)
- reaffirm
- reiterate
- report

- say
- show
- state
- stress, emphasize, underscore
- suggest
- tell
- propose
- point out
- request

The following verbs as candidates for **cognitive acts**:

- anticipate
- being concerned
- believe
- calculate
- conjecture
- dream
- expect

- estimate
- fear
- feel like (in the sense of *think*) (taken from RST-DT))
- figure out
- hope
- imagine

- know
- predict
- recognize
- suppose
- think
- understand
- wish

We list the following verbs which we did **not** regard as cognitive nor speech acts:

- agree (because it is an action, and not a cognitive act)
- ask (in the sense of "want" [3])
- being certain

- being clear
- being proud
- confess
- demonstrate
- ensure
- thank (for + inf)

- make sure
- note (in sense of *noticing*))
- realize
- urge (to + inf)

The verb *ask* is a speech act, but not if used in a sense of *want* like in the next example:

> *We ask that all measures be taken to facilitate the work of the investigators on the ground.* (UNSC_2014_SPV.7219_spch011_sentsplit_France.txt)

Another question left open in both RST-DT and GUM is (possibly because it is typical for the UNSCon speeches but not for the texts they annotated), if **non-factual attributions** should be still marked as attribution, like calls from one party to another party to make an speech act, or hypothetical speech acts. We decided to also include those cases, if the (hypothetical) reported speech or cognitive act is more central then the speech/cognitive act itself.

> [We also call upon Russia,] [as a signatory to the Vienna document,] [to help secure their release] **[and to confirm publicly,]** [even if belatedly,] [for the record] [that the ab-

ducted monitors were part of a legitimate mission on behalf of the international community.]
(UNSC_2014_SPV.7165_spch008_sentsplit_United_States_Of_America.txt)

2.2.6.3 Complements of Nouns and Adjectives

Noun complements, including prepositional phrases, that-clauses, and to-infinitive complements, help specify the noun's content, characteristics, or purpose, enhancing the overall meaning and clarity of the sentence. Noun complements are normally used with abstract nouns which represent an idea rather than a thing. The complement, or additional information, explains what that idea relates to. For example:

[Unanimous decisions and acceptance of whatever is put on the table is something that longs to regimes, not democratic Governments.] UNSC_2014_SPV.7165_spch016_sentsplit_Lithuania.txt

[The scenario we are seeing reminds one of the events that took place in Crimea one month ago:] (...) (UNSC_2014_SPV.7154_spch008_sentsplit_France.txt)

[The **possibility** that the actions of some violent groups could threaten the holding of this democratic exercise is unacceptable.] (UNSC_2014_SPV.7165_spch005_sentsplit_France.txt)

In the first example above, *acceptance* is an abstract noun complemented by the phrase of *whatever is put on the table*. Combined, the noun and complement make one complete, grammatical idea. As we stated in Rule X, we will not segment the noun complement. The coordinate *not democratic Governments* is not segmented since we state that that an proposition is only segmented if it is complete and not an object clause, see 2.2.7. In the second example, *events* is the abstract noun which demands more specification (the could also be easily replaces with the demonstrative pronoun those, see Rule X in section 2.2.2.2), although this is a borderline case (since theoretically the sentence "The scenario reminds us of the events" would be possible, too, but not probable). The this example is also a borderline case, showing the difficulty in segmenting relative clauses and deciding if they are noun complements and therefore not segmented or relative clauses, and therefore segmented.

The frequently used extra-clausal constituent in English: *the (DET) (adj) fact (is) (that)* is not a sufficient marker for segmenting before *that*. Only if it would be segmented even if leaving out the constituent, we segment it. The first sentence could be reformulated in "... we are deeply concerned that the Geneva agreement of 17 April is not yet *achieving* its objective *to restore* stability in Ukraine." The first part is an attribution followed by an infinitiv construction, which would not be segmented. Therefore we also do not segment it. It is also not segmented when it is only a filler like in the second example, where the dash can be replaced with *This is*, which would be *This is a fact that has been affirmed [...]* and therefore not segmented.

[The United Kingdom requested this meeting] [because we are deeply concerned by the deteriorating security situation in southern and eastern Ukraine] [and by *the fact that* the Geneva agreement of 17 April is not yet achieving its objective to restore stability in Ukraine.] UNSC_2014_SPV.7165_spch004_sentsplit_United_Kingdom_Of_Great_Britain_And_Northern_Ireland.txt

[Russian nationals are not under threat in Ukraine] [- *a fact that* has been reaffirmed on numerous occasions, including by the United Nations Human Rights Monitoring Mission and by the Commissioner for National Minorities of the Organization for Security and Cooperation in Europe (OSCE).] (same speech)

2.2.6.4   Infinitival Complements versus Purpose Clauses

As RST-DT stated, it is sometimes not straightforward to distinguish between infinitival complement (not segmented) and purpose clauses (which are segmented and can be tested by substituting the phrase *in order to* for the *to* clause). Also, a purpose clause does modify the entire main clause and not only a noun (see RST-DT, p. 64).

Two examples from the UNSCon:

> [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.] (UNSC_2014_SPV.7138_spch002_sentsplit_Russian_Federation.txt)

> [Like others] [who preceded me] [have done,] [Belgium launches an appeal to all the parties] [*to guarantee full access to the disaster site,* not only for the search and rescue teams] [but also for the investigators.] (UNSC_2014_SPV.7219_spch029_sentsplit_Belgium.txt)

The first EDU in the first example above is not segmented, since the preposition *to* is complementing the adjective *important* and cannot be replaced by *in order to* but only *for*. In the relatively complex second example, *who preceded me* is an embedded relative clause splitting up the clause *like others (…) have done,*. The sense of the clause *to guarantee full access […]* would not be modified if *to* would be substituted with *in order to*, therefore it is a purpose clause and is segmented. The last EDU is segmented since it is part of a strong discourse marker *not only…but also*, see section 2.2.6. For more examples how to distinguish between infinitival complements and purpose clauses we refer to the RST-DT guidelines Carlson et al. (2002), section 2.3.1, p. 6.

2.2.7   Coordinate Clauses and Elliptical Constructions

2.2.7.1   Coordinate Clauses

> Rule X: Full clauses coordinated by 'and', 'or' or 'but' are made into independent EDUs.

As already stated, one clause states usually one EDU. The simplest way to combine clauses in English grammar is by using a coordinating conjunction like *and*, *but*, *so* or *or* (*I will go shopping and you can bring Marry to school.*). The coordinate clause is thereby a clause that is introduced by one of the coordinating conjunctions. Usually the segmentation between coordinated clauses is straightforward (*[I will go shopping] [and you can bring Marry to school.]*). Generally, the coordinating conjunction (here: *and*) belongs to the second EDU.

2.2.7.2   Elliptical Constructions

Nevertheless, coordination can create elliptical constructions, in which a word or phrase that is implied by the context is left out of a sentence, usually because it is a repetition of a preceding word or phrase. As Stede et al. (2017) already state, handling of elliptical constructions is sometimes tricky. In the following, we will discuss some rules for those cases.

From GUM we adopt the following rules for elliptic constructions:

Rule X: An elliptic coordination **is segmented**, if:

(1) they share the main verb (i.e., are coordinate Verb Phrases). Ex. from GUM: *[Mary drank coffee] [and Jane tea].*

(2) the subject is shared, but not the object (see next rule). Ex. from GUM: *[The harbor master called] [and sorted us out]* (*the harbor master* is the subject ob both verbs).

(3) modals and auxiliaries are shared. This does not prevent EDU segmentation for coordinated infinitives (see rule (2) for not segmented elliptical constructions). Ex. from GUM:*[Kim should call first] [and then go]*

Rule X: An elliptic coordination **is not segmented**, if:

(1) a coordination is inside an object clause. Ex. from GUM: *[They were ordered to call the harbor master and get sorted out]*, where *to call the harbor master* is the object clause and *to* is shared across the elliptic clause (*what were they ordered to do?*).

(2)) the object is shared across both clauses and no EDU intervenes, the sentence is considered a single EDU. Ex. from GUM: *[The harbor master ordered and prepared the boat]* (boat is a shared object (what did they ordered? and what did they prepared?), no segmentation of coordination)

(3) a coordinate VP is the to-infinitive object of another verb. Ex.: *[The company plans to shut down its plant and dismiss several hundred employees.] [He wanted to go and check]* This is also following RST-DT, who state that infinitival complements are never EDUs, also not in elliptic constructions (RST-DT, p.13).

Following the first rule, the object clauses in the next example starting with *in accordance with (…)* are therefore not segmented:

> [That principle is confirmed in the 1970 Declaration on Principles of International Law] [concerning Friendly Relations and Co-operation among States in accordance with the Charter of the United Nations, and in a number of other decisions of the General Assembly, as well as the Helsinki Final Act of 1975.] (UNSC_2014_SPV.7138_spch002_sentsplit_Russian_Federation.txt)

In the next sentence, there is a relative clause *the principle of equal rights and self-determination of peoples* which left out in the beginning *which is the principle (…)*, there it is not segmented.

> [The philosophy of the sponsors of the draft resolution runs counter to one of basic principles of international law, the principle of equal rights and self-determination of peoples,] [enshrined in Article 1 of the Charter of the United Nations.] (UNSC_2014_SPV.7138_spch002_sentsplit_Russian_Federation.txt)

Coordinated complements of Attributes on the other hand are segmented when they would normally be segmented as an EDU if it were a single clause:

> [They must stop claiming] [that we have deployed our army along the border with that country and are practically already at the English Channel,] [and that we have sent agents] [to coordinate the protests in Ukraine.] (UNSC_2014_SPV.7138_spch002_sentsplit_Russian_Federation.txt)

But if the complement is not segmented as attribution it is not segmented, like in the next example, where there is no source of attribution and therefore it is no legitimate candidate for segmenting:

[It must be understood that south-eastern Ukraine and its people are deeply worried about their future and that they do not want anyone, especially the nationalist radicals, to impose their will on them.] (UNSC_2014_SPV.7138_spch002_sentsplit_Russian_Federation.txt)

We included an exception for the next example, where the complements are segmented via semi-colons, pointing towards an list:

[The Council does not have to take my word for it] [- or even that of the Ukrainian Government -] [it need only witness for itself the videos of professional military shepherding thugs into a building in Kramatorsk;] [the photographs showing the so-called concerned citizens taking over Sloviansk, equipped exactly like the elite troops that took Crimea;] [or the video of a military operation in Krasny Liman by armed men with the same equipment.] (UNSC_2014_SPV.7154_spch006_sentsplit_United_States_Of_America.txt)

With these examples we want to show that still some decisions were made not only based on syntactic rules but decided case by case also due to the content of the elliptic complements.

There are cases where the object clauses are themselves a chained structure, and can include for ex. relative clauses. When strictly implementing our rules we potentially loose discourse information for the sake of consistency. Since we can distinguish between ELABORATION (provides details or more information on the *general* state of affairs described in the EDU) and E-ELABORATION (providing more information on a *single* entity) (see XX) we decided to enable relative clause segmentation for unsegmented elliptic coordinations, using E-ELABORATION as relation label.

[We call on Russia commit to de-escalation and to condemn (...) the attempts at destabilization] *[carried out by armed groups in eastern Ukraine.]* (UNSC_2014_SPV.7154_spch008_sentsplit_France.txt)

For more examples for each case, please refer to Zeldes (2017). Additionally we include the rule taken from RST-DT:

> Rule X:If the main verb is ellipted, we do not segment, only if there is a strong discourse marker (see section 2.2.6).

Also, if coordinate clauses appear in embedded units, such as relative clauses or postmodifiers, the clauses are segmented by the usual rules of clausal segmentation. (ex. [She signed up,] [starting as an "inside" adjuster,] **[who settles minor claims]** **[and does a lot of work by phone.]**) (From RST-DT, sec. 2.9.6., p. 23:)

Stede et al. (2017) define as a rule for dividing an elliptical construction in case there are clear hints for a coherence relation to be present. In the next example, the enumeration of propositions serves a such a hint. Another criterion for segmentation that PCC-RST states is that a proposition is by and large complete, that is why the last EDU is not segmented.

[Nobody knows why she does not come,] [who takes her place,] [when or whether she will come back.] (maz-5876, PCC), example taken from (Stede et al., 2017).

As we stated, the splitting of coordinations in the UNSCon sometimes is not straightforward, also since the speakers in the Council sometimes tend to use complex constructions. We want to exemplify this in the next example.

[China will continue to adopt a fair and objective approach and actively participate in any proposal or initiative] [that will reduce tension and promote a political solution.] (UNSC_2014_SPV.7219_spch008_sentsplit_China.txt)

In the example it is not unambiguously when the elliptic construction starts. We can read it as *China will continue to adopt a fair and objective approach | China will continue to actively participate in any proposal or initiative.*, then the object clause is shared among the coordinated clauses, therefore it is not segmented. We could also read it as *China will continue to adopt a fair and objective approach. | China will actively participate in any proposal or initiative.* Nevertheless here we would not segment it because of RuleX that infinitival complements are not segmented.

In the next example, *to unequivocally establish the facts behind the incident* is not segmented because of the infinitival construction, but *and to find out exactly what happened and who is responsible.* is segmented.

> [We owe it to the victims of this tragedy and their families to unequivocally establish the facts behind the incident] [and to find out exactly what happened and who is responsible.] (UNSC_2014_SPV.7219_spch009_sentsplit_Australia.txt)

2.2.8 Prepositional Phrases

Prepositional clauses are not a necessary signalling a discourse segmentation. Following GUM, we state that they are only segmented if they have a verb as head.

Identifying prepositions is sometimes not as straightforward as it seems. Next to the typical prepositions of time or place, there also exist **participle prepositions**. Some verbs form participles (ending with *-ing*, *-en*, or -ed) that can act as an preposition. Some of the common participle prepositions in English are: *excluding, including, following, considering, regarding, notwithstanding, given.* Additionally, we found in the corpus often appearing: with regard to, in terms of. The next sentence is therefore not segmented:

> [*Following* close to two weeks of relative calm, *starting* on 6 April armed individuals and separatist groups began to seize Government buildings and confiscate lethal weapons in Donetsk oblast of Ukraine, *including* in the cities of Donetsk, Lugansk and Kharkiv.] (UNSC_2014_SPV.7154_spch002_sentsplit_UN.txt)

As we stated, prepositional phrases containing *with regard to* are only segmented when containing a verb. While the first example is segmented, the second example could be seen as a borderline case, since "upholding" is a verbal normalization of the verb "to uphold", nevertheless, we do not segment it since it is a noun phrase.

> [It is imperative to protect the rights of the citizens of Ukraine] [regardless of ethnic background.] (UNSC_2014_SPV.7154_spch009_sentsplit_Rwanda.txt)

> [The Kyiv Government made certain promises with regard to constitutional reforms and upholding the rights of minorities.] (UNSC_2014_SPV.7165_spch015_sentsplit_Russian_Federation.txt)

Next to regarding/with regard to, RST-DT defines *regardless (of)* as special type of strong discourse marker which is be segmented even when not containing a verb (see RST-DT, sec. 2.10., p. 26 and section 2.2.6.1). We differ here slightly from the PCC-RST guidelines who demand an verb to be an EDU (see RST-PV, 3.2.4, p.6-7 compared to RST-DT, 2.10, p.26).

Another potentially less straightforward issue is deciding between gerunds and normalization of verbs after a preposition, like in the next example where addressing is nominalized and therefore not segmented:

> [The fundamental *importance of addressing* the impact of conflict on women and girls is now widely acknowledged.] (UNSC_2016_SPV.7658_spch017_sentsplit_New_Zealand.txt)

2.2.9 Segmentation for References to Documents or Meetings

The speeches in the corpus often refer to other speeches or documents like resolutions including a speech-ID in parenthesis in the texts. We treat those references roughly as citations in GUM and RST-DT, where citations and dates in parenthetical are usually segment. Only when they are syntactically integrated in the sentence they are not segmented.

> [However, we voted for the draft resolution contained in document S/2014/189 because it reasserts the primacy of the principle of territorial integrity.] (UNSC_2014_SPV.7138_spch011_sentsplit_Argentina.txt) – *not segmented*

> [Egypt is proposing an amendment to paragraph 2 of the draft resolution] [(8/2016/235)] [that has been submitted to the Security Council.] [On Thursday, we heard a clear message from the Prime Minister of Ukraine] [that Ukraine was willing to engage in dialogue with Russia] [to address its stated concerns] [(see S/PV.7134).] (UNSC_2016_SPV.7643_spch002_sentsplit_Egypt.txt)

The reference in the last example (*(see S/PV.7134).*) is connected via the Evidence relation to the two preceding EDUs, similar to GUM, who choose explanation-evidence. For more information on the labels used for references, see section 3.3.

2.2.10 Difficult Cases: Relative Clauses

(1) **Dashes:** In the next example, *[...] - disbelief at the situation that developed before our eyes* is not segmented because we read the dash as standing for "it is", which would therefore be a cleft-construction: "It is a disbelief at a situation that developed before our eyes". Also, we did not segmented it because the second part is necessary for the meaning (it could be easily interpreted as a cleft: "It is a situation that developed before our eyes.").

> [It is with a sense of disbelief that we find ourselves here today] [- disbelief at the situation that developed before our eyes] [with the Russian Federation's determination to annex Crimea.]
>
> (UNSC_2014_SPV.7138_spch005_sentsplit_France.txt)

Similarly, the phrase *of those who* is not segmented and treated as cleft construction:

> [Safe access must also be provided] [to allow fro the recovery of the bodies] [of those who lost their lives in the tragedy.]

(2) In the next example (which is a shortened sentence from the UNSCon corpus) the relative clause *that were predicted* could have been segmented according to our rules as defined in section 2.2.2. Nevertheless we see it as cleft construction as it could have been changed into *There has been no trace of the hundreds of thousands of predicted refugees*. It could also be seen as an Attributive complement clause, but since there is no source, it is not segmented.

> [There has been no trace of the hundreds of thousands of refugees that were predicted.] (UNSC_2014_SPV.7138_spch005_sentsplit_France.txt)

> [The individuals involved called for secession from Ukraine] [and for joining the Russian Federation.] (UNSC_2014_SPV.7154_spch002_sentsplit_UN.txt)

Similarly, the second example could be reformulated into The involved individuals called (...). Even if in principle we could segment *involved* 2.2.2 since it states an embedded zero-relative clause splitting up the clause, it is only referring to the noun phrase *the individuals*. We decided to avoid such fine-grained segmentation.

2.2.11 Other Remarks

(1) The adverb *namely* is included at the beginning of the second EDU when it is segmented, as in the relative clause in the next example:

> [They all point in one direction,] [*namely,* that this horrible crime was committed as part of the armed aggression against Ukraine.] (UNSC_2014_SPV.7219_spch007_sentsplit_Lithuania.txt)

(3) Following GUM, **fillers** like *you know* or *I mean* or *see* are segmented using XX relation, as they prepare the other EDU without contributing content to the discourse.

There are also sentences that do not have a verb or subject, as required by RST-DT and GUM for an EDU. That is also because speeches – although in large parts prewritten – sometimes also are given in free speech. Like fillers, we segment those sentences as one EDU and use XX relation label.

> [Basically, it is simple:] [the Russian veto today is telling us that might is right.] *[Well, no.]* [Force cannot override the law.] (UNSC_2014_SPV.7138_spch005_sentsplit_France.txt)

(3) The next prepositional phrase with a subject drop *as VERB+ed* is segmented even if there is no subject and therefore no clause:

> [Lithuania urges the holding of a full, thorough and independent investigation into the downing of the Malaysian Airlines flight,] *[as agreed today by the Council.]* (UNSC_2014_SPV.7219_spch007_sentsplit_Lithuania.txt)

(4) Adnominal clauses are only separated if it is followed by a verb clause. In the next example, the first and last EDU is connected via an SAME-UNIT clause

> [The news] [that the lives of 80 children were lost] [is devastating.] (UNSC_2014_SPV.7219_spch009_sentsplit_Australia.txt)

(5) Discourse markers like *however*, *furthermore*, or *moreover* are generally not segmented. If they are embedded between attributive and main clause they are integrated into the EDU of the former:

> [I think,] [**however,** that my Russian colleague and I will find common ground in a key sentence from the speech of the Soviet Permanent Representative in 1976 on the question of Mayotte,] [who said:] [...]
>
> (UNSC_2014_SPV.7138_spch005_sentsplit_France.txt)
>
> [However, we believe] [it is important to underscore once again that we cannot accept attempts to exclude national peacekeeping contingents...] (UNSC_2016_SPV.7643_spch007_sentsplit_Russian_Federation.txt)

(6) **Colons:** Following RST-DT, text after colons is treated as separate EDUs as long as they provide further elaboration on the topic introduced. Those EDUs doesn't need to be clauses (first example) but can be phrases (second example) or even one word. One exception is, that if the colon is inside an noun phrase, it won't be segmented.

> [Since yesterday, we have seen the abrupt and simultaneous outbreak of violence in eastern Ukraine:] [nothing about it was spontaneous.]

[The resolution allows us to send a clear message to all actors:] <– Elaboration[Blue Helmets, police, civilian personnel and international forces.] (UNSC_2016_SPV.7643_spch009_sentsplit_France.txt)

[We are striving to elaborate an effective plan] [to address the four pillars of the agenda:] [prevention, protection, participation, and peacebuilding and recovery.] (UNSC_2016_SPV.7658_spch029_sentsplit_Angola.txt)

Sometimes dashes can have the same function as colons, therefore the same segmentation is applied.

[We must act on what we stand for] [- peace and security.] (UNSC_2014_SPV.7154_spch009_sentsplit_Rwanda.txt)

(7) In the next example We interpret "in the face of" as being another formulation for "facing", therefore we segmented the prepositional clause.

[We can only imagine what any other Government would have done] [in the face of such provocation on its national territory.] (UNSC_2014_SPV.7165_spch005_sentsplit_France.txt)

2.3   Content-based Segmentation: Utterances of Conflicts

From RST-DT we take the rule that relative clauses are not segmented if they include only an adjective (-able) without a verbal element. However, we allow the exception, if the content of the unit is strongly pointing to a conflict (as defined in the UNSCon paper), we can segment it.

Although relative clauses are strictly defined as containing a verbal element, we allow some interpretation and exceptions, based on as to whether or not it has a function to express a conflict, like in the next example. Here the reference to the chamber is important for interpreting a potential attack or critique on members of the chamber. While this is a content-based decision that may be different for other corpora and tasks, we decided it was important not to not loose those conflict-segments and to analyze them rhetorical framework.

[Some people,] [including in this Chamber,] [do not wish to recognize the real reasons for the events in Ukraine and are forever seeing the hand of Moscow in the South-East.] (UNSC_2014_SPV.7154_spch004_sentsplit_Russian_Federation.txt)

3 Rhetorical Relations in the UNSCon

3.1 Relation Set in UNSCon

We follow the labels as defined in Stede et al. (2017), but additionally include four labels taken from RST-DT, which are explained in the following.

3.1.1 Same-Unit (multinuclear, NN)

We include SAME-UNIT, a pseudo-relation used as a device for linking two discontinuous text fragments that are really a single EDU, but which are broken up by an embedded unit. Examples of embedded units that can break up other EDUs include: relative clauses, other nominal postmodifiers, parentheticals, participial clauses, etc. By convention, this relation is always multinuclear.

> [But maintaining the key components of his strategy] [– **a stable exchange rate and high levels of imports** –] [will consume enormous amounts of foreign exchange.]

> [United Nations peacekeeping operations should,] [**within the purview of their respective mandates,**] [assist host countries] [in safeguarding the security of women in conflict] [and should effectively implement a zero-tolerance policy toward sexual violence.] (UNSC_2016_SPV.7658_spch015_sentsplit_China.txt)

We allow units to be interrupted by complex embedded unit, meaning more than one EDU can break up a legitimate EDU. In the next example the EDUs in bold interrupt the legitimate EDU *Undertakings must be upheld.*

> [Undertakings] [**brokered by the trilateral contact group with armed separatists**] [**who control the territory in which the crash occurred**] [**to allow access**] [must be upheld.] (UNSC_2014_SPV.7219_spch009_sentsplit_Australia.txt)

For more info on the segmentation of EDUs please refer to section 2.2.4.

Adjacent units which state one EDU but are segmented due to sentence segmentation are also connected. For those interrupted EDUs based on false sentence-splitting, we also use SAME-UNIT.[4]

3.1.2 Attribution (mononuclear, SN)

Normally, clausal complements of verbs are not considered to be EDUs. Adapting the ATTRIBUTION relation from RST-DT, we make an exception in the case of clausal complements of attribution verbs, including both speech acts and other cognitive acts. The attribution predicate is the satellite and the attributed material, also called reported message, is the nucleus. Excluded from the annotation are infinitival complements from attribution relations, attribution predicates that do not identify a source and passive constructions.

For verbs of attribution that occur in the middle of a stream of reported speech, select an embedded attribution relation if the source breaks up a single EDU, as in Example (1). Select a non-embedded attribution relation if the source occurs between two separate EDUs, as in (2). In both cases, the attribution satellite is linked to the first segment of the split EDU.

(1) ["Seeing Michelle up there,"] [she added,] ["was like watching myself or my daughter."]

(2) ["When Sears has a sale at a special price,"] [the woman in the ad declares,] ["it's something you don't want to miss."]

---

[4]Also refer to section 2.1.2 and section 2.1.3 for more information how we handle incorrectly parsed documents for the EDU segmentation in the UNSCon corpus

Speech acts – verbs that are used to report both direct and indirect speech – should be segmented (examples: say, tell, state, announce, declare, suggest, advise, report, indicate, point out, explain, ask). They are marked for the rhetorical relation of ATTRIBUTION, if there is an explicit source of the attribution. We mark the attribution (the clause containing the reporting verb) as the satellite, and the content of the reported message (which must be in a separate clause) as the nucleus.

Cognitive predicates, including verbs that express feelings, thoughts, hopes, etc., should also be segmented and marked for the rhetorical relation of ATTRIBUTION (examples: think, believe, know, imagine, suppose, conjecture, wish, hope, predict, fear, estimate, calculate, anticipate, expect, dream).

3.1.3 Textual-Organization (mononuclear, SN)

TEXTUAL-ORGANIZATION is a relation used to link elements of the structure of the text, for example, to link a title with the body of the text, a section title with the text of a section, etc. The role of the relation is primarily that of enforcing a tree structure on the representation. We follow GUM in that a heading can be seen as a preparation text for the body, where the first is the nucleus and the former the satellite. Therefore in those cases TEXTUAL-ORGANIZATION is used mononuclear. Summarizing, we use TEXTUAL-ORGANIZATION for:

- section headings (mononuclear);
- included meta-info from the speech (like the language the speaker speaks) (mononuclear);
- and parts of speeches from other speakers (mononuclear).

3.1.4 Topic-Comment (mononuclear, SN)

We use TOPIC-COMMENT for EDUs not contributing propositional content to the discourse, including backchanneling, incomplete or repaired/aborted utterances, and fillers. For those cases, we follow GUM that the interrupting sequence is the satellite of the main EDU.

TOPIC-COMMENT is also used for relating questions and their answers, where one textual span poses a question (not necessarily realized as an interrogative sentence), and the other text span answers the question. For those cases, N is the answer to the question posed by S. We also use TOPIC-COMMENT for rhetorical questions, where no answer is expected but used to emphasize a point made. Here the related EDU poses a question vis-a-vis a segment of the text. And finally, TOPIC-COMMENT is used when one segment is a statement and the other span makes some sort of response to it.[5]

For more theoretical thoughts on questions and rhetorical questions handled in RST, please refer to the next section 3.2.1.

Summarizing, we use TOPIC-COMMENT (all mononuclear) for:

- backchanneling, incomplete or repaired/aborted utterances. Ex.: [I mean,]$_{\text{TOPIC-COMMENT}}$ [I had to do it].; [There was no other option,] [you know.]$_{\text{TOPIC-COMMENT}}$
- question-answers
- when one segment is a statement and the other span makes some sort of response to it.

3.1.5 Join Justification and Reason

Because of minor semantic differences (they both want to change the attitude of the reader) and to reduce the size of labels, we merge JUSTIFICATION and REASON labels and keep the latter.

---

[5]In RST-DT, TOPIC-COMMENT always states a multinuclear relation, as usually both spans are necessary to understand the context (p. 70). Nevertheless, we follow GUM in this stating that a hierarchical nucleus-satellite relation can be established for the mentioned cases.

3.2   Further Considerations on RST Labeling in UNSCon

3.2.1   (Rhetorical) Questions handled in UNSCon

We decided to label questions with the TOPIC-COMMENT relation **when they do not provide any additional information to the content.**[6] For those cases we adopt the TOPIC-COMMENT class for questions from RST-DT, which includes the labels question-answer (similar to GUM's topic-question label). In GUM, when a question is a filler-formluation (like "You know?") they are marked as *organization-phatic*, which is mapped to RST-DT again with *Topic-Comment* Class in the mapping made in Liu et al. (2023), which we follow.

Nevertheless, the speeches also include rhetorical questions used to emphasize a particular idea or persuade the audience of a point, giving an answer within the question raised. Since rhetorical questions often have the purpose to emphasise for example a REASON for a claim, or the EVALUATION of a situation or statement, etc., we decided to use the appropriate label, instead of a general TOPIC-COMMENT class. We only use TOPIC-COMMENT, if it would be possible to exclude the question without loosing much of the content.

In RST-DT, Rhetorical questions are labeled as *Rhetorical-Question*, which is a sub-type of TOPIC-COMMENT. Nevertheless ideally, rhetorical labels should express the purpose it has for another discourse unit and not just be a description of the text segment itself. Additionally, RST-DT only defines that the *Rhetorical-Question* is a satellite, but it was not clear to us how to detect the nucleus since there is no answer to be related.

3.3   References

As mentioned in section 2.2.9, the speeches in the corpus often refer to other speeches or documents like resolutions including a speech-ID in parenthesis in the texts. We treat those references roughly as citations in GUM and RST-DT, where citations and dates in parenthetical are usually segment. Only when they are syntactically integrated in the sentence they are not segmented. Following GUM, references forming an EDU (i.e. non-syntactically integrated, see segmentation guidelines) typically function as EVIDENCE:

> [That was evident at last week's open debate on conflict prevention in the Great Lakes region] [(see S/PV.7653).] EVIDENCE

> [On Thursday, we heard a clear message from the Prime Minister of Ukraine] [that Ukraine was willing to engage in dialogue with Russia] [to address its stated concerns] [(see S/PV.7134).]EVIDENCE (UNSC_2016_SPV.7643_spch002_sentsplit_Egypt.txt)

Other possible relation for references is CIRCUMSTANCE, when they include dates describing events, years, etc. (similar to GUM, section *Parentheticals and references*).

> [Russia backed the adoption of resolution 2272] [(2016),]CIRCUMSTANCE [on sexual exploitation and abuse in peacekeeping operations.] (UNSC_2016_SPV.7643_spch007_sentsplit_Russian_Federation.txt)

3.4   Acronyms and Translations

Following GUM, if an acronym for an expression within a sentence is specified in parentheses, it is considered a satellite partial restatement, and marked with RESTATEMENT relation.

> [The way that peacekeeping troops and troop-contributing countries] [(TCCs)]RESTATEMENT [have been libelled is completely unacceptable.] (UNSC_2016_SPV.7643_spch005_sentsplit_Egypt.txt)

---

[6]Questions and rhetorical questions are not mentioned in P-V RST guidelines but appear in the UNSCon speeches.

4 Relations used for UNSCon and Mapping with other frameworks

1

Table 1 lists all labels used for UNSCon and maps them according to the Corresponding GUM V8 and RST-DT Classes, taken from Liu & Zeldes (2023). We use → if a relation can only point forward by definition; ← if only backward; →← if either forward or backward; NN for multi-nuclear and NS for nucleus-satellite relations.

GUM V8 Relations	GUM V8 Classes	Corresponding RST-DT Classes	UNSCon RST relations
adversative-antithesis	Adversative	Contrast	Antithesis (NS, →←)
adversative-concession	Adversative	Contrast	Concession (NS, →←)
adversative-contrast	Adversative	Contrast	Contrast (NN, →←)
attribution-positive	Attribution	Attribution	Attribution (NS, →←)
attribution-negative	Attribution	Attribution	Attribution (NS, →←)
causal-cause	Causal	Cause	Cause (NS, →←)
causal-result	Causal	Cause	Result (NS, →←)
context-background	Context	Background	Background (NS, →←)
context-circumstance	Context	Background	Circumstance (NS, →←)
contingency-condition	Contingency	Condition	Condition (NS, →←)
contingency-condition	Contingency	Condition	Unless (NS, →←)
contingency-condition	Contingency	Condition	Otherwise (NS, →←)
elaboration-attribute	Elaboration	Elaboration	Elaboration-E (NS, →←)
elaboration-additional	Elaboration	Elaboration	Elaboration (NS, →←)
explanation-evidence	Explanation	Explanation	Evidence (NS, →←)
~~explanation-justify~~	~~Explanation~~	~~Explanation~~	~~Justify (NS, →←)~~
explanation-justify	Explanation	Explanation	Reason-S (NS, →←)
explanation-justify	Explanation	Explanation	Reason-N (NS, →←)
explanation-motivation	Explanation	Explanation	Motivation (NS, →←)
evaluation-comment	Evaluation	Evaluation	Evaluation-S (NS, →←)
evaluation-comment	Evaluation	Evaluation	Evaluation-N (NS, →←)
evaluation-comment	Evaluation	Evaluation	Interpretation (NS, →←)
joint-disjunction	Joint	Joint	Joint (NN)
joint-list	Joint	Joint	List (NN)
joint-conjunction	Joint	Joint	Conjunction (NN)
joint-sequence	Joint	Temporal	Sequence (NN)
joint-other	Joint	Topic-Change	Joint
mode-manner	Mode	Manner-Means	Means (NS, →←)
mode-means	Mode	Manner-Means	Means (NS, →←)
organization-phatic	Organization	Topic-Comment	Topic-Comment (NS, →←)
organization-preparation	Organization	Textual-Organization	Preparation (NS, →)
organization-heading	Organization	Textual-Organization	Textual-Organization (NS, →←)
purpose-attribute	Purpose	Enablement	Purpose (NS, →←)
purpose-goal	Purpose	Enablement	Enablement (NS, →←)
restatement-partial	Restatement	Summary	Summary (NS, →←)
restatement-repetition	Restatement	Summary	Restatement (NS, →←)
topic-question	Topic	Topic-Comment	Topic-Comment (NS, →←)
topic-solutionhood	Topic	Topic-Comment	Solutionhood (NS, →←)
same-unit	same-unit	Same-Unit	Same-Unit (NN)

Table 1: Relation Mapping for UNSCon between different RST frameworks, mapping between GUM's GUM and RST-DT is taken from Liu & Zeldes (2023).

Acknowledgments

Use unnumbered third level headings for the acknowledgments.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. RST Discourse Treebank. 2002. URL https://catalog.ldc.upenn.edu/LDC2002T07.

J.R. Hurford. Grammar: A Student's Guide. Cambridge University Press, 1994. ISBN 9780521456272. URL https://books.google.com.af/books?id=ZaBKd8pT6kgC.

Yang Janet Liu and Amir Zeldes. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In Andreas Vlachos and Isabelle Augenstein (eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3112–3130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.227. URL https://aclanthology.org/2023.eacl-main.227.

Yang Janet Liu, Tatsuya Aoyama, and Amir Zeldes. What's hard in english rst parsing? predictive models for error analysis. arXiv preprint arXiv:2309.04940, 2023.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. Text - Interdisciplinary Journal for the Study of Discourse, 8(3): 243–281, 1988. doi: doi:10.1515/text.1.1988.8.3.243. URL https://doi.org/10.1515/text.1.1988.8.3.243.

Mirco Schönfeld, Steffen Eckhard, Ronny Patz, and Hilde van Meegdenburg. The UN security council debates 1995-2017, 2019. URL http://arxiv.org/abs/1906.10969.

Manfred Stede, Maite Taboada, and Debopam Das. Annotation guidelines for rhetorical structure. 2017. URL https://www.sfu.ca/~mtaboada/docs/research/RST_Annotation_Guidelines.pdf.

Amir Zeldes. The GUM corpus: Creating multilayer resources in the classroom. Language Resources and Evaluation, 51(3):581–612, 2017. doi: http://dx.doi.org/10.1007/s10579-016-9343-x.

A   The first appendix

This is an example of an appendix.

B   The second appendix

This is an example of a second appendix. If there is only a single section in the appendix, you may simply call it "Appendix" as follows:

Appendix

This format should only be used if there is a single appendix (unlike in this document).