Annotator disagreement in RST annotation schemes

Daniil Ignatev[1], Denis Paperno[1], Massimo Poesio[1,2],

[1]Utrecht University, [2]Queen Mary University of London,

Correspondence: d.ignatev@uu.nl

Abstract

Discourse parsing within the Rhetorical Structure Theory (RST) framework has inspired extensive research; however, it remains prone to significant levels of annotator disagreement, particularly in the labeling of relations and nuclearity. This paper investigates systematic discrepancies in RST annotations, focusing on two expert-annotated corpora of closely related languages. We first compare different RST treebanks to assess the availability of parallel-labeled data and highlight their usefulness for studying disagreement. We then perform both quantitative and qualitative analyses of annotation divergences, identifying factors that contribute significantly to inconsistent interpretations. Finally, we propose two practical approaches for addressing disagreement: (1) filtering out unhelpful biases and (2) capturing legitimate ambiguity through more flexible annotation schemes.

1 Introduction

In the field of computational linguistics, discourse parsing — particularly within the Rhetorical Structure Theory (RST) framework — offers a well-established approach to analyzing the coherence relations between different parts of a text. This task involves identifying and classifying discourse relations, such as the cause-effect relationship, between individual units, like sentences or paragraphs. Foundational work by Mann and Thompson (Mann and Thompson, 1988) and advancements by Daniel Marcu (Marcu, 1996, 2000) have introduced methodologies for constructing trees that represent discourse units and their connections, ultimately reflecting the rhetorical composition of texts. In RST, elementary discourse units (DUs) are roughly analogous to clauses, but higher order units can span indefinitely up to a complete text. The framework employs 30 relations to capture the full range of connections between these units. Related spans are classified into nucleus and satellite, where the nucleus represents the central or more significant unit of the relation[1].

The complexity inherent in discourse annotation frequently leads to disagreements among annotators at multiple levels. Even rigorously designed RST corpora, such as RST-DT (Lynn Carlson, 2002), the Potsdam Commentary Corpus (Stede and Neumann, 2014), and the Dutch Discourse Treebank (van der Vliet et al., 2011; Redeker et al., 2012), typically yield kappa scores reflecting at best substantial agreement.

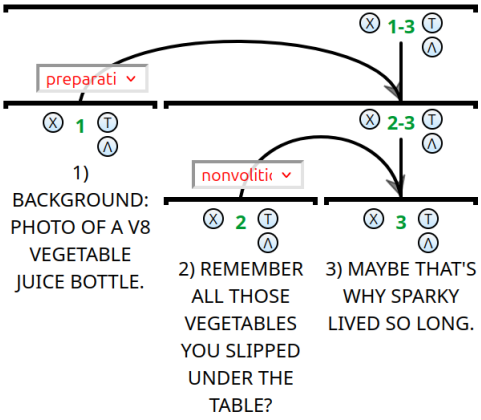

Figure 1: Example from RST website (Taboada and Mann, 2006) in RSTWeb (Zeldes, 2016). Cropped labels: preparation, nonvolitional cause

On the other hand, while the subject of disagreement in discourse annotation has been widely addressed in theory, there have been relatively few suggestions on how this issue could be addressed in practice. Meanwhile, recent years have seen

[1]Beyond RST, other frameworks such as the Penn Discourse Treebank (PDTB, Prasad et al. 2008) and Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003) have explored alternative approaches to labeling discourse relations. For the former, there exists a body of work dealing with disagreement (Yung et al., 2024; Scholman and Demberg, 2017), showing that this problem is relevant for either framework.

the emergence of a large body of work on learning from disagreement, proposing a number of approaches to handling varying interpretations. In natural language processing, transformer-based architectures (Devlin et al., 2019; Liu et al., 2019) are increasingly used to capture nuanced linguistic phenomena, and this includes work on leveraging label distributions and annotator-specific biases (Rodrigues and Pereira, 2017; Mostafazadeh Davani et al., 2022). Such strategies include augmenting the gold standard based on the spectrum of opinions (Plank et al., 2014; Fornaciari et al., 2021), learning from distributions of labels using a soft metric (Sheng et al., 2008; Aroyo and Welty, 2014; Peterson et al., 2019; Uma et al., 2020), and training separate models on labels coming from individual annotators (Akhtar et al., 2020). However, despite these trends, deeper engagement with disagreements in discourse-level tasks like RST parsing has been limited.

Given this tendency, addressing the research gap mentioned above becomes increasingly important. To this end, we pursue several objectives in this paper:

- Review existing RST resources with respect to the extent of disagreeing annotations they contain.
- Perform quantitative and qualitative analyses of factors contributing to disagreement, using suitable data sources.
- Based on the obtained results, propose preferable ways of integrating disagreements into RST annotation and RST parsing.

The scope of this paper primarily concerns RST relations and nuclearity, leaving aside two other major aspects of RST: segmentation of text into EDUs and organizing these segments into spans. While these areas are also subject to disagreement and require thorough analysis, we exclude them here for several reasons. Firstly, in most existing corpora, inter-annotator agreement on these tasks is much higher compared to relation and nuclearity labeling (see Das et al. 2017 for details). Additionally, in most flavors of RST annotation, EDU segmentation is grounded in syntax and leaves considerably less room for subjective interpretation. This is evident to the extent that some RST parsers assume text segmentation is given; while debatable, this assumption remains widely adopted in practical applications (Maekawa et al., 2024).

Our results suggest that RST annotation is substantially influenced by individual preferences of annotators, which sometimes conflict with the annotation manual. In such cases, considering the entire range of disagreeing annotations seems redundant. On the other hand, a larger portion of disagreements is prompted by factors that allow for multiple interpretations, making the adoption of a spectrum of readings by individual experts a generally feasible strategy.

2 Related Work

2.1 Theories of disagreements in discourse annotation

The subject of discrepancies in RST analysis has been widely discussed in the community, with particular attention given to the relational level.

In this context, two notions need to be distinguished: first, one annotator assigning multiple complementary relations; second, several annotators assigning multiple relations that may or may not be complementary. We will refer to the former as "multi-level" annotation and the latter as "disagreement." While our primary focus is on the latter, the concept of multi-level analysis suggests that diverging concurrent analyses may all be plausible: if one annotator can assign multiple complementary relations to the same span, it is reasonable to assume that several annotators can do the same. For this reason, we consider the respective arguments in the discussion, even though they do not concern disagreement directly.

(1) The topic of multi-level analysis has been widely discussed in the literature. For example, Mann and Thompson, 1988 suggested that multiple relations can be assigned to the same span. Similarly, Moore and Pollack, 1992 argued that each relation between rhetorical units should be annotated on two levels: informational and intentional, as the existing relation types exhibit significant overlap with respect to these domains. Arguments in favor of multi-level annotation have since appeared in numerous works (see Taboada and Mann, 2006 for a systematic overview).

However, Sanders and Spooren, 1999, followed by Stede, 2008a, oppose this suggestion, claiming that complex annotation would be redundant in most cases, as the relations involved are typically either exclusively informational or exclusively intentional. Meanwhile, Taboada and Mann, 2006 notes that postulating multiple relations may be jus-

tified in ambiguous cases that cannot be resolved based on the context.

(2) The issue of ambiguity in the RST framework has been directly addressed in several works by Manfred Stede (Dipper and Stede, 2006; Stede, 2008b,a), based on the experience of building the Potsdam Commentary Corpus (Stede and Neumann, 2014). The results of this work are summarized in Stede, 2008a, which identifies several sources of ambiguity in RST annotations, such as vagueness in definitions and conflicting scopes of relations, and argues that many of these can be resolved through distinguishing several levels of discourse annotation: thematic, referential, and others. To that end, the work introduces a specialized framework, MLA.

A related line of work (Iruskieta et al., 2015; Wan et al., 2019) proposed changes to how the similarity of structures should be measured in RST annotations. The alternative metrics penalize discrepancies on different levels (relation directionality, nuclearity, relation type) differently, depending on how important each factor is for the overall structure.

Finally, some recent works suggest a permissive approach to concurrent interpretations, advocating for their incorporation into the gold standard. (Das et al., 2017) compare amateur and expert RST annotations in English and German and propose treating competing expert analyses as a "complex ground truth." They suggest Underspecified Rhetorical Markup Language (URML, Reitter and Stede, 2003) as a means of storing discourse graphs. On the other hand, eRST, a proposal for RST enhancement, allows for additional edges, i.e., concurrent relations, in RST structures, provided these relations are realized lexically through discourse markers. Although this notion does not directly address disagreements, it enables the integration of several alternative analyses into one structure and permits at least some alternative readings on the relational level. In other words, parallel annotations in existing corpora can partially be integrated into eRST graphs.

2.2 Analyzing Annotation Discrepancies

Qualitative analyses of disagreements have primarily been conducted by corpus designers. For instance, da Cunha et al., 2011 examined disagreements in Spanish RST. A significant amount of qualitative analysis of RST disagreements, which ultimately remained unpublished, was carried out by the authors of the Dutch Discourse Treebank (NLDT) based on their own material. While we conducted our qualitative analysis independently on a subset of their corpus, resulting in different hypotheses, we extend our gratitude to Gisela Redeker for granting us access to their data and observations (Redeker and van der Vliet, 2015).

3 Datasets with disagreements

Given the known complexities and disagreements in RST annotations, it has become standard practice in corpus design to include at least a small subset of texts annotated independently by multiple annotators, facilitating measurement of inter-annotator agreement. However, there are substantial differences in how many documents receive parallel annotations, how many discourse units these documents include, and how many annotators are involved. These differences have implications for how helpful the annotations are for learning from disagreement: although the amount of suitable data remains the most important factor, it is certainly not the only one.

Despite this common practice, some datasets lack parallel annotations. Specifically, the Georgetown University Multilayer Corpus (Zeldes, 2017), currently the largest RST treebank, used a development procedure that purposefully avoids measuring the relative annotation quality; as a result, the corpus does not have parallel markup[2]. The Basque RST treebank did not have parallel annotations on the level of whole documents, as its developers measured disagreement on granular tasks, such as the assignment of causal relations (Iruskieta et al., 2013); aside from that, only reconciled annotations are available in the public release. For several corpora, there exist a number of parallel annotations, but these have not been made publicly available for various reasons. This applies to the Potsdam Commentary Corpus (Stede and Neumann, 2014) and APA RST (Hewett, 2023).

Some resources are offered by the RST Discourse Treebank (Lynn Carlson, 2002), formerly the largest RST dataset, containing 385 newswire texts from the Wall Street Journal section of the Penn Treebank. Fifty-three texts from this main corpus body received parallel annotations, providing a relatively large set of parallel RST structures that was published with the main corpus. Still, some

[2]Secondary edges from eRST graphs cannot be fully considered as such, since, for instance, they are not independent from primary ones.

Corpus	N annotators	N docs	N EDUs	Notes
Dutch RST	3	80	2344	Docs unevenly split: 80 / 74 / 13
Kobalt RST	2	42	2216	
CSTNews 6.0	2	5	97	3 or 4 versions for some docs.
Russan RST	3	3	225	
APA RST	3	36	-	*Non-public
RST DT	-	52	2938	*Non-attributed
Spanish RST	-	80	694	*Non-attributed

Table 1: Parallel data in RST corpora. N EDUs assumes the gold standard segmentation.

factors limit the utility of this data for analyzing disagreement.

- Firstly, the primary corpus annotations are not independent of the parallel annotations, as the former result from a reconciliation process involving these parallel versions.
- Secondly, annotations are not explicitly attributed to individual experts, limiting the analysis of annotator-specific perspectives or biases.

The Spanish RST treebank shares the latter two issues, although it remains one of the largest sources in terms of parallel texts, comprising around 700 discourse segments distributed across 80 parallel documents.

For a number of RST treebanks, the opposite is true, i.e., the data is attributed and produced by workers independently, but its amount is insufficient to conduct a feasible quantitative analysis. Such is the case with the Brazilian (CSTNews 6.0, Cardoso et al., 2011) and Russian treebanks (Toldova et al., 2017). We provide the number of annotated documents for these and other corpora in Table 1.

Finally, several corpora feature substantial amounts of attributed parallel annotations, though these are not publicly available and must be requested directly from their creators. A notable example is the Dutch Discourse Treebank (NLDT), which offers three annotation versions for each of its 80 documents (comprising 2,344 EDUs). Typically, two experts annotated each text independently (with a third annotator occasionally participating), followed by a reconciled version (van der Vliet et al., 2011; Redeker et al., 2012). For our analysis, we selected 74 texts annotated by the two experts responsible for the largest annotation share. Although the annotations are not anonymized, for the purpose of our study, we treat the annotators anonymously, labeling them experts A, B, and C.

Another corpus with the desired properties is Kobalt RST (Wan, 2021), a subset of the Kobalt corpus annotated with discourse trees. Similarly, its 42 documents (comprising 2216 EDUs) have three versions: two readings by experts and a reconciliation. Although Kobalt covers a very specific genre of discourse, i.e., argumentative essays by non-native German speakers, it remains suitable for analyzing RST disagreements, such as eliciting individual biases of annotators. We do not incorporate the reconciled annotations in our experiments, as we aim to preserve the raw perspective of each annotator.

Remarkably, both Kobalt and NLDT were annotated by trained experts holding at least a master's degree in linguistics or related disciplines. This expertise level (see Das et al. 2017) and their higher motivation as opposed to crowd annotators ensure the quality of their work. Another similarity is that Kobalt and NLDT concern related languages allowing for a cross-language comparison (which, however, has to account for lexical and syntactic differences). These similarities are another reason why we use both Kobalt and NLDT in our further analysis.

4 An analysis of disagreements in the datasets

In this section, we compare disagreements across the two corpora more closely by reporting confusion matrices and inspecting the label pairs where annotators show consistent divergences. To avoid dealing with matrices that are too nuanced and sparse, we only accounted for cases of disagreement on relations and disregarded cases where ex-

perts agree on a relation but disagree on nuclearity[3]. We pay special attention to whether the experts' markup exhibits systematic disagreements. To that end, we consider the most frequently confused relations, dividing them into two categories: "symmetrical" cases, in which annotators A and B confuse relations X and Y equally frequently or at least similarly often, and "asymmetrical" cases, where confusing X and Y is only typical for annotator A or B.

In the first category, we note several tendencies: firstly, problematic relation pairs often involve the ELABORATION relation. Although the annotation manuals for Kobalt and NLDT, the former based on PCC (Stede and Neumann, 2014), treat it differently, it still remains a frequent option that experts resort to when unable to assign a more precise label. While the notion of this relation being problematic has been around for a long time, it is even more evident in a cross-lingual comparison on attributed material. Of more interest is that CAUSE in Kobalt is often confused with other relations by both annotators, sometimes multinuclear and non-causal (LIST). Inspecting the data instances manually, we notice that 81% of these are lexically unspecified and involve adjacent sentences, as in (1).

(1) [Überregionale Produkte werden so stark wie nie konsumiert .] $\xrightarrow{\text{CAUSE/LIST}}$ [Die heutige Generation profitiert von einem vielfältigen Warenangebot dank der Globalisierung ... .]$_{\text{Kobalt\_DEU\_004}}$

Understandably, in this setting, experts struggle to agree on the relative importance of sentences, since normal heuristics, like the deletion test[4], are harder to apply. Likewise, the causality of the relation is also debatable, as human opinions on whether one statement entails another can diverge greatly, as shown by other text understanding tasks (Nie et al., 2020). Some other prominent disagreements, such as those involving JUSTIFY and MOTIVATION in NLDT (Redeker and van der Vliet, 2015), also occur in this underspecified setting.

We report the most frequent disagreements from the second category in Table 2 & Table 3. One of the tendencies we find remarkable is the great number of disagreements over multinuclear relations. This could offer insight into the high value of length as a feature, since multinuclear relations, especially JOINT, which can be used to link arbitrary parts of text, tend to occur in an intersentential position and thus their respective spans are longer in length. Based on the provided numbers, it can be argued that annotators tend to develop a preferred reading for ambiguous cases and assign a specific label based on past experiences. Such is the case with NONVOLITIONAL CAUSE from NLDT, which annotator A considers applicable to a wider range of situations: overall, in our subset of corpus data, expert A uses NONVOLITIONAL CAUSE 111 times, while expert B only 86. Other relations with a similar skew are BACKGROUND (40 vs. 23), JOINT (57 vs. 27), and, to a lesser extent, CONJUNCTION (231 vs. 262).

Incidentally, some of the confusions we observe in Kobalt are also characteristic of other RST corpora: da Cunha et al., 2011 report CONCESSION and ANTITHESIS to be frequently confused in the Spanish treebank. On the other hand, unlike Spanish RST, MEANS and CIRCUMSTANCE are almost never confused in the two corpora, suggesting that the authors' explanation based on connective polysemy is correct.

Relations	Ann. A	Ann. B
conjunction-list	**26**	2
joint-list	2	**7**
concession-antithesis	**8**	0

Table 2: Frequent preferences in Kobalt

Relations	Ann. A	Ann. B
joint-conjunction	1	**22**
nonvol-cause-nonvol-res	**12**	3
list-joint	**11**	1
summary-preparation	**8**	1
nonvol-cause-circumstance	**7**	1

Table 3: Frequent preferences in NLDT

4.1 Results: discussion

The previously made observations shed some light on how various cases of disagreement are distributed in the corpora; we argue that a significant part of these does not constitute an informative signal. One example of this is ELABORATION: keeping this label as an alternative to more specific relations may not be particularly helpful for

[3] We report the most frequently confused relations in the appendix in Table 5 & Table 6.

[4] The deletion test involves removing each part of a relation in turn to determine whether the entire span would retain its original meaning. The part that is harder to delete is considered more important.

understanding the text by either human or machine readers, since the more specific relation often implies that one discourse unit elaborates on the other. Preserving ELABORATION may also have undesired effects during parser training, as parsers tend to develop a bias towards it as the most frequent relation. A further example is constituted by relation types that experts subjectively prefer — possibly, contrary to annotation rules. For instance, the confusion between CONJUNCTION and LIST observed in Kobalt may be a case of this, as the respective manual suggests that LIST should only be assigned when lexical or graphic signals explicitly indicate an enumeration. In cases like that, only one annotator is "correct" with respect to the manual.

However, there also remain plausible divergences in the analyses that can prove informative if preserved in the annotation, such as the CAUSE/LIST example above. The factors behind cases like that include both conflicting or ambiguous signals (several DMs etc.) and underspecification; the latter leads to conflicting readings especially frequently (as another example, consider MOTIVATION and JUSTIFY in NLDT).

In order to determine the more suitable strategy for preserving the meaningful disagreements, it is essential to consider the relative impact of these factors. In the following section, we propose a computational experiment for that purpose.

5 Modeling disagreements

5.1 Motivation

Our experiment aims to quantify the relative impact of surface variables on annotator disagreement, particularly, on discourse relations. In order to do so, we train a classifier for a binary objective: whether two annotators agree or disagree on the relation class given two related discourse units. Our assumption is that signals that consistently prompt diverging interpretations will emerge as important features, while irrelevant signals will not make an impact. To that end, we pick XGBoost as a classifier model that can leverage feature combinations and robustly estimate their contribution (Chen and Guestrin, 2016). As an example, Liu et al., 2023 and Pastor and Oostdijk, 2024 both used XGBoost to analyze hard and easy signals in RST parsing. We also consulted both of these works when determining the set of features.

5.2 Enhancing datasets

For our experiments, we ensured that both corpora were annotated for relevant syntactic and discourse variables, such as UD tags and discourse markers. This required additional intermediate steps as described below.

Concerning syntactic features, we addressed the problem of dependency tagset mismatch. For NLDT, syntactic dependency markup using the Universal Dependencies (UD) standard was published in 2023 as part of the DisRPT shared task (Braud et al., 2023). In contrast, the dependency annotations available for Kobalt use the Hamburg Dependency Treebank (HDT, Borges Völker et al., 2019) annotation standard, which, aside from different tags, also displays a number of differences in tree-building rules (Shadrova, 2020). To ensure that both of our models used syntactic features of similar granularity, we converted the existing dependency annotations from the HDT to the UD standard using a robust converter developed by (Hennig and Köhn, 2017) and obtained standard CONLL-U files.

Discourse features presented a different challenge, namely, the need for a uniform way of annotating both datasets with discourse markers. The task of detecting and disambiguating discourse connectives has drawn significant attention in the context of PDTB-style discourse parsing, with several tools developed specifically for these tasks (Dipper and Stede, 2006; Bourgonje and Stede, 2020). However, these tools only target German and lack a Dutch counterpart. Another development in this direction is the creation of discourse connective inventories for both languages: DimLex (Stede and Umbach, 2002) and DisCoDict (Bourgonje et al., 2018), in which all entries are additionally annotated for possible non-connective readings.

In our approach, we leveraged natural language instructions and used OpenAI's text-to-text generative model O1-mini (OpenAI, 2023) to highlight DM candidates. We purposefully based the model's instructions on a relaxed definition of discourse markers (compared to PDTB), synthesized from Fraser, 2009's account. Our motivation was to cover the entirety of discourse marker candidates to assess their impact on experimental results. The respective prompts are provided in Section A in the appendix.

In the absence of gold DM annotations, we tested the efficiency of this solution using a rule-based

baseline that, while imperfect on its own, provides a reliable approximation of ground truth. Specifically, this baseline highlights all entries from DimLex or DisCoDict in the text using regular expressions; however, we discard all matches except those that occur at an EDU-initial position (assuming the existing EDU segmentation). This choice is based on the understanding that a large portion of DM candidates, such as “und” or “en” (“and”) or “als” (“when”), occur at the start of a clausal EDU when acting as subordinating conjunctions and, consequently, as discourse connectives.

We then tested O1-mini’s robustness in detecting these EDU-initial DM candidates, resulting in accuracy scores of 79% and 83% on Kobalt and NLDT, respectively. This, along with a manual inspection we conducted, demonstrates that both O1-mini’s predictions and the baseline show reasonable reliability.

Regarding sources of errors, we note that a large portion of misclassifications occurs due to GPT selecting markers that do not fall into the definition of a discourse connective in PDTB terms and are thus absent from the lexicons we used. These alleged false positives include instances such as “gelukkig” (“luckily”) or “overigens” (“besides”); whether these can truly be regarded as connectives remains an open question.

5.3 Predicting disagreements

Similarly to Liu et al., 2023 and Pastor and Oostdijk, 2024, we do not train the classification algorithm on the text of the two discourse units but only supply it with pre-extracted features. Originally, the features we use were found to be related to item difficulty and could, thus, help predict disagreements; we supply the full list below:

- Discourse unit length in symbols;
- Number of discourse markers (`dm_count`), type of the head DM, i.e., a DM that is the highest in the constituent hierarchy of the second span (`dm`);
- Dependency function of a discourse unit’s syntactic head (`DEPREL` of the head in CONLL-U terms);
- Number of elementary discourse units (roughly, number of clauses) in the first and the second discourse unit, and in total;
- Genre, when applicable;
- Intra-, inter- (involving two sentences), or multisentential status of the relation (Redeker and van der Vliet, 2014) as three binary features;
- Lastly, the label assigned by one of the two annotators, which helps understand whether the experts are in two minds over some particular relation types.

As in the parser-oriented study (Liu et al., 2023), we split our features into two groups. The first group comprises surface features that experts can utilize when annotating a text, while the second group includes the full set of features. The surface feature group includes the following attributes: DU length, the number and type of discourse markers, the syntactic function of the head, and the inter-, intra-, or multisentential status.

Dataset	All	Surface
Kobalt	0.75	0.73
NLDT	0.68	0.59

Table 4: Mean F1 score of XGBoost (5-fold CV)

For each dataset, we separately utilize two subsets of features: surface-only features ("realistic") and all features. We report the average F1 score across a 5-fold cross-validation in Table 4 and provide the relative weights for all factors in Figure 2. It can be seen that, in general, the classifier does not attain an optimal score, especially on NLDT, where the model based on surface features performs slightly above chance. This may indicate that the collected features are insufficient or, at least, do not correlate well with disagreement in NLDT.

5.4 Results: discussion

Despite different classification scores, the two models exhibit a clear pattern in terms of the features they select as relevant. Concretely, discourse unit length always emerges as the most important factor. When the "label" feature is included, it is always the next deciding factor, suggesting that annotators consistently disagree over specific relations: e.g., one picks CAUSE while another picks EXPLANATION. Lastly, the head’s syntactic function and DM type also make a contribution in all settings, although their role in Kobalt seems to be more prominent. Importantly, DM variable appears not as informative as other factors[5].

[5]Evidence from PDTB annotation also demonstrates that agreement does not hinge on the presence of markers: inter-

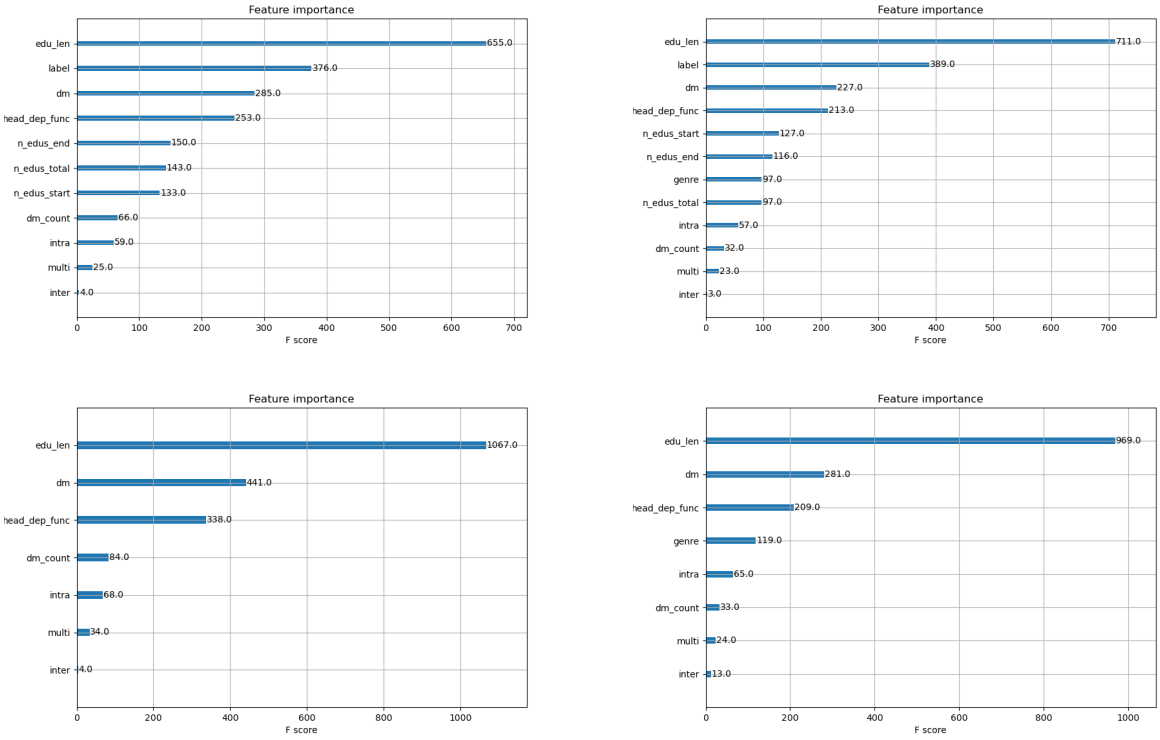

Figure 2: XGBoost weights for Kobalt (left) and NLDT (right): all (top) and surface-only (bottom) features. Abbreviations: edu_len: discourse unit length; dm: discourse marker (present/absent); dm_count: discourse marker count; intra/inter/multi: intra-, inter-, or multisentential relation span; n_edus_start/end: number of EDUs within the first/second argument of the relation.

The first notion aligns well with some of the existing hypotheses about automatic discourse parsing, namely, that humans and parsers struggle more when analyzing relations between lengthy spans of text, as in Nguyen et al., 2021; Shi et al., 2020. Nevertheless, unit length proves to be consistently more important than similar features that account for syntax or tree position: longer spans are often multisentential and include more elementary discourse units, but these factors do not emerge as important.

6 Discussion

The results of our analysis allow us to speculate about the best way of preserving meaningful RST interpretations. As mentioned in Section 2, the two existing alternatives are URML (Das et al., 2017) and eRST (Zeldes et al., 2024); the former of these two could incorporate all parallel readings, and the latter only those that are lexically grounded, i.e., based on one or two discourse markers. Here, we would like to address two properties of eRST annotation that make it less feasible for this task.

The first of these is its definition of discourse markers, which serve as a basis for secondary edges. In this respect, eRST aligns completely with PDTB's notion of discourse connectives and its respective restrictions: only subordinating conjunctions, coordinating conjunctions, and adverbials can have the status of discourse markers (Zeldes et al., 2024). In this paper, we are not looking to contribute to the vast theoretical discussion on what lexical elements should be considered discourse markers; however, we must note that existing studies offer different answers to this question, sometimes using the same linguistic material. For instance, annotating the Wall Street Journal corpus with PDTB-style discourse connectives (PDTB 2.0, Prasad et al., 2008) and with more vaguely defined discourse markers (RST Signalling Corpus, Das and Taboada, 2017; Das, 2014) results in a different number of unique markers being identified: 100 and 201, respectively. Partly, this is due to the latter category including combinations like "but also", but also due to inclusion of broader lexical

annotator agreement for implicit relations (85.1%, Prasad et al., 2008) is only slightly lower than for explicit ones (90.2%).

categories.

Undoubtedly, adopting a stricter definition simplifies the task for corpus annotators, resulting in better reliability of their work. On the other hand, it raises the question of whether using a broader set of markers, such as that of the RST Signalling Corpus, would allow for broader coverage of secondary edges and better reflect the space of possible interpretations of discourse—something that eRST, as well as ourselves, seeks to address. For example, such items as "naturally", "of course", and "after all" are not listed as explicit in either PDTB 2.0 (Prasad et al., 2008) or PDTB 3.0 (Prasad et al., 2019). However, we could model cases where "naturally" would signal REASON relation and "after all" would signal CAUSE. In eRST terms, it would prompt the addition of a primary or a secondary edge.

(2) [We only left home at 8;] $\xleftarrow{\text{REASON}}$ [naturally, we were late.]

(3) [He will do that for you,] $\xrightarrow{\text{CAUSE}}$ [because, after all, he is your brother.]

These examples suggest that relaxing the existing lexical criteria for secondary edges could, in theory, improve coverage.

A further possible shortcoming of eRST is that it cannot incorporate plausible readings of underspecified relations unlike URML. This is especially important since in the existing corpora, the larger part of relations is not signalled by markers (Taboada, 2006; Das and Taboada, 2017). Our observations also confirm that disagreement is strongly associated with underspecification; thus, we argue that a standard that aims to integrate parallel readings will profit from allowing multiple graph edges in underspecified cases.

7 Conclusion

The analyses presented in this paper highlight that RST annotations exhibit a persistent and systematic degree of inter-annotator disagreement. Drawing on two expert-annotated corpora (Dutch and German), we observe that divergent interpretations often arise from the inherent complexity of discourse relations, especially when label definitions are underspecified or conflated. Although some discrepancies reflect an annotator's systematic bias (e.g., favoring ELABORATION or LIST), in many cases, multiple readings of a relation are equally plausible. Our experiments suggest that span length and certain label choices serve as strong predictors of disagreement, indicating that large or complex discourse spans are particularly prone to ambiguous interpretations.

From an applied perspective, two complementary strategies emerge. First, filtering out demonstrable biases that run counter to annotation rules can clarify the "true" consensus. Here, the judgment needs to be based around surface signals handled differently than prescribed; consequently, even rule-based systems or simpler neural language models can prove helpful at this task.

Second, adopting flexible schemes that capture legitimate ambiguity, such as URML or eRST, can more comprehensively reflect discourse complexity; of these two, we find URML better suited for this (and only for this) specific task, as it gives more freedom for genuine discrepancies to be integrated. Moving forward, these dual approaches — tightening clearly defined guidelines while embracing multiple valid analyses — hold promise for improving both the reliability and the expressive power of RST annotation.

Limitations

We acknowledge that our analysis focuses on RST relations paying less attention to the partly overlapping problems of disagreements in nuclearity and discourse unit spans. Furthermore, we highlight that the features we used when predicting disagreement do not offer an exhaustive picture of factors behind annotation discrepancies. Considering additional variables, such as rhetorical "moves" (Redeker et al., 2012) or syntactic signals beyond clause boundaries, could make the analysis more complete.

Acknowledgments

We thank the anonymous reviewers, Hugh Mee Wong, Frances Yung, and the NLP group at the University of Utrecht for their helpful feedback on this work. We are thankful to Gisela Redeker and Gosse Bouma for granting us access to the early versions of the Dutch discourse treebank. This work is funded by NWO through an AINed Fellowship Grant NGF.1607.22.002 and supported by project 'Dealing with Meaning Variation in NLP'.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opin-

ions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Hum. Comput.*, 1:31–44.

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Peter Bourgonje, Jet Hoek, Jacqueline Evers-Vermeul, Gisela Redeker, Ted J M Sanders, and Manfred Stede. 2018. Constructing a lexicon of dutch discourse connectives.

Peter Bourgonje and Manfred Stede. 2020. Exploiting a lexical resource for discourse connective disambiguation in german. In *International Conference on Computational Linguistics*.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Paula Christina Figueira Cardoso, Erick Galani Maziero, Mara Elena Lucia, R. Castro Jorge, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças, Volpe Nunes, Thiago Alexandre Salgueiro Pardo, and Rodovia Washington Luís. 2011. Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the rst spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Debopam Das. 2014. *Signalling of Coherence Relations in Discourse*. Ph.D. thesis, Simon Fraser University.

Debopam Das, Manfred Stede, and Maite Taboada. 2017. The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 11–19, Santiago de Compostela, Spain. Association for Computational Linguistics.

Debopam Das and Maite Taboada. 2017. Rst signalling corpus: a corpus of signals of coherence relations. *Language Resources and Evaluation*, 52:149 – 184.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Stefanie Dipper and Manfred Stede. 2006. Disambiguating potential connectives.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Bruce L. Fraser. 2009. An account of discourse markers. *International Review of Pragmatics*, 1:293–320.

Felix Hennig and Arne Köhn. 2017. Dependency tree transformation with tree transducers. In *UDW@NoDaLiDa*.

Freya Hewett. 2023. APA-RST: A text simplification corpus with RST annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 173–179, Toronto, Canada. Association for Computational Linguistics.

Mikel Iruskieta, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The rst basque treebank : an online search interface to check rhetorical relations.

Mikel Iruskieta, Iria da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, 49(2):263–309.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Mary Ellen Okurowski Lynn Carlson, Daniel Marcu. 2002. *RST Discourse Treebank*. Philadelphia: Linguistic Data Consortium.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in rst discourse parsing by using large language models? *Preprint*, arXiv:2403.05065.

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.

Daniel Marcu. 1996. Building up rhetorical structure trees. In *Proceedings of AAAI-96*, pages 1069–1074, Portland, OR.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Johanna D. Moore and Martha E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. RST parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Martial Pastor and Nelleke Oostdijk. 2024. Signals as features: Predicting error/success in rhetorical structure parsing. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 139–148, St. Julians, Malta. Association for Computational Linguistics.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *CoRR*, abs/1908.07086.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. Abacus Data Network.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multilayer discourse annotation of a dutch text corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association (ELRA).

Gisela Redeker and Nynke van der Vliet. 2014. Explicit and implicit coherence relations in dutch texts. *Pragmatics and beyond. New series*, 254:23–52.

Gisela Redeker and Nynke van der Vliet. 2015. Exploring and evaluating rst annotations. *Unpublished manuscript*.

David Reitter and Manfred Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*.

Filipe Rodrigues and Francisco Pereira. 2017. Deep learning from crowds. *Preprint*, arXiv:1709.01779.

T.J.M. Sanders and W.P.M.S. Spooren. 1999. *Communicative intentions and coherence relations*, pages 235–250. Number 63 in Pragmaticsamp;Beyond. J. Benjamins.

Merel C. J. Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *LAW@ACL*.

Anna Valer'evna Shadrova. 2020. *Measuring coselectional constraint in learner corpora: A graph-based approach*. Ph.D. thesis, Humboldt-UniversitÃ¤t zu Berlin, Sprach- und literaturwissenschaftliche FakultÃ¤t.

Victor S. Sheng, Foster J. Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. *Organizations & Markets eJournal*.

Ke Shi, Zhengyuan Liu, and Nancy F. Chen. 2020. An end-to-end document-level neural discourse parser exploiting multi-granularity representations. *CoRR*, abs/2012.11169.

Manfred Stede. 2008a. Disambiguating rhetorical structure. *Research on Language and Computation*, 6:311–332.

Manfred Stede. 2008b. Rst revisited : disentangling nuclearity.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Manfred Stede and Carla Umbach. 2002. Dimlex: A lexicon of discourse markers for text generation and understanding. In *International Conference on Computational Linguistics*.

Maite Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592. Focus-on Issue: The Pragmatics of Discourse Management.

Maite Taboada and William C. Mann. 2006. Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, 8:423 – 459.

Svetlana Toldova, Dina Pisarevskaya, M. I. Ananyeva, Maria Kobozeva, Alexandr Nasedkin, S. Nikiforova, Irina Petrovna Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in russian rst treebank.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *AAAI Conference on Human Computation & Crowdsourcing*.

N.H. van der Vliet, I. Berzlánovich, G. Bouma, M. Egg, and G. Redeker. 2011. Building a discourse-annotated dutch text corpus. In *Beyond Semantics*, volume 3 of *Bochumer Linguistische Arbeitsberichte*, pages 157 – 171. Ruhr-Universität Bochum.

Shujun Wan. 2021. Kobalt_rst (rst german learner treebank): die annotation von rhetorischen strukturen im kobalt-daf-korpus.

Shujun Wan, Tino Kutschbach, Anke Lüdeling, and Manfred Stede. 2019. RST-tace a tool for automatic comparison and evaluation of RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 88–96, Minneapolis, MN. Association for Computational Linguistics.

Frances Yung, Merel C. J. Scholman, Sárka Zikánová, and Vera Demberg. 2024. Discogem 2.0: A parallel corpus of english, german, french and czech implicit discourse relations. In *International Conference on Language Resources and Evaluation*.

Amir Zeldes. 2016. rstWeb - a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, San Diego, California. Association for Computational Linguistics.

Amir Zeldes. 2017. The gum corpus: creating multilayer resources in the classroom. *Lang. Resour. Eval.*, 51(3):581–612.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2024. erst: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, pages 1–47.

A Connective detection prompts

A.1 NLDT connective detection prompt

Instruction
In the following Dutch text, identify all discourse markers (DMs) and enclose them in <dm> tags.
Definition of Discourse Markers (DMs):
- DMs, also known as connectives, are lexical expressions (e.g., *en*, *maar*, *omdat*, *dus*, *hoewel*, *toch*) that belong to different syntactic classes such as conjunctions, adverbials, and prepositional phrases.
- They are used to connect discourse components (text segments) and signal the coherence relations that hold between those components (e.g., contrast, cause, elaboration).
- The scope of a DM's function is a single discourse sequence comprising adjacent text spans in a relation.
- DMs can be present at the beginning, middle, or end of a sentence (or segment).
- A DM signals relations that hold between two adjacent text segments but does not create the relation; it guides the interpretation of the relation.
Guidelines:
1. **Scope of DMs:**
- The function of a discourse marker applies to a single discourse sequence comprising adjacent text spans in a relation.
- DMs signal relations that hold between two adjacent text segments.
- A discourse marker does not create the relation between text segments; it only guides the interpretation of the relation.
2. **Position of DMs:**
- DMs can be present at the beginning, middle, or end of a sentence (or segment).
- They may appear within the sentence or at clause boundaries.
3. **Identification of DMs:**
- Use a list of common Dutch DMs to identify potential markers, such as:
- **Addition:** *en*, *ook*, *bovendien*
- **Contrast:** *maar*, *echter*, *toch*

- **Condition:** *als*, *indien*, *tenzij*

- **Cause/Reason:** *omdat*, *want*, *doordat*

- **Concession:** *hoewel*, *ofschoon*, *desondanks*

- **Temporal:** *toen*, *terwijl*, *voordat*, *nadat*

- **Result/Consequence:** *dus*, *daardoor*, *zodat*

- **Example:** *bijvoorbeeld*, *zoals*

- Ensure the word functions as a DM in context by connecting two propositions or clauses.

- Confirm that the token's part of speech corresponds to typical DM categories (conjunctions, adverbials, prepositional phrases).

4. **Annotation Format:**

- Enclose each identified DM within <dm> and </dm> tags.

- Do not alter the original text other than adding the tags around the DMs.

5. **Examples:**

English Example:

Input:

"A country is considered financially healthy **if** its reserves cover three months of its imports."

Output:

"A country is considered financially healthy <dm>if</dm> its reserves cover three months of its imports."

Dutch Examples:

Example 1:

Input:

"Drie nieuwe emissies beginnen vandaag te handelen op de New York Stock Exchange, **en** één begon vorige week te handelen op de Nasdaq/National Market System."

Output:

"Drie nieuwe emissies beginnen vandaag te handelen op de New York Stock Exchange, <dm>en</dm> één begon vorige week te handelen op de Nasdaq/National Market System."

Example 2:

Input:

"De Poolse rat zal deze winter goed eten. Tonnen heerlijk rottende aardappelen, gerst en tarwe zullen vochtige schuren over het hele land vullen **terwijl** duizenden boeren de kopers van de staat wegsturen."

Output:

"De Poolse rat zal deze winter goed eten. Tonnen heerlijk rottende aardappelen, gerst en tarwe zullen vochtige schuren over het hele land vullen <dm>terwijl</dm> duizenden boeren de kopers van de staat wegsturen."

Task:

- Read the following Dutch text.

- Identify all discourse markers based on the guidelines above.

- Enclose each DM within <dm> tags.

- Ensure that the rest of the text remains unchanged.

Notes:

- Pay special attention to words that can function as DMs but may have other grammatical roles. Use context to determine their function.

- The goal is to produce a text identical to the input except for the addition of <dm> tags around the identified discourse markers.

- Do not tag words that are not functioning as discourse markers in the given context.

By following these instructions, you will identify and annotate all discourse markers in the text, which will help in analyzing the coherence relations within the text and assist in computational processing.

Text to Process:

A.2 Kobalt connective detection prompt

Instruction

In the following German text, identify all discourse markers (DMs) and enclose them in <dm> tags.

Definition of Discourse Markers (DMs):

- DMs, also known as connectives, are lexical expressions (e.g., und, weil, obwohl) that belong to different syntactic classes such as conjunctions, adverbials, and prepositional phrases.

- They are used to connect discourse components (text segments) and signal the coherence relations that hold between those components (e.g., contrast, cause, elaboration).

- The scope of a DM's function is a single discourse sequence comprising adjacent text spans in a relation.

- DMs can be present at the beginning, middle, or end of a sentence (or segment).

- A DM signals relations that hold between two adjacent text segments but does not create the relation; it guides the interpretation of the relation.

Guidelines:

1. **Identify Potential DMs:**

- **Common DMs in German include:**

- **Conjunctions:** und (and), aber (but), oder (or), denn (for), sondern (but rather), weil (because), obwohl (although), wenn (if), während (while), falls (in case).

- **Adverbials:** deshalb (therefore), trotzdem (nevertheless), allerdings (however), außerdem (besides), folglich (consequently), inzwischen (meanwhile), dennoch (still).

- **Prepositional Phrases:** im Gegensatz zu (in contrast to), aufgrund von (due to), trotz (despite), infolgedessen (as a result).

2. **Position in Sentence:**

- DMs can appear at the beginning, middle, or end of a sentence.

- Examples:

- Initial: <dm>Trotzdem</dm> geht er zur Arbeit. (Nevertheless, he goes to work.)

- Medial: Er geht <dm>trotzdem</dm> zur Arbeit.

- Final: Er geht zur Arbeit, <dm>trotzdem</dm>.

3. **Confirm the Function:**

- Ensure the word or phrase is functioning as a DM and not in another grammatical role.

- Exclude words that are not functioning as DMs (e.g., "dass" as a complementizer). Exclude "dass" as a complementizer. Exclude "dass" as a complementizer. Exclude "dass" as a complementizer.

- Exclude "dass" as a complementizer.

- Exclude "und" if not interclausal.

Examples:

1. **Example (English DMs):**

- **Relation DMs:**

- Circumstance: when, as, with

- Condition: if, unless

- Contrast: but, however

- Concession: while, though

- Elaboration-additional: and, also

- Reason: because, due to

- List: and, in addition, moreover

- Temporal-after: since, after

- Temporal-before: before

2. **Example 1:**

Three new issues begin trading on the New York Stock Exchange today, <dm>and</dm> one began trading on the Nasdaq/National Market System last week. On the Big Board, Crawford & Co., Atlanta, (CFD) begins trading today. Crawford evaluates health care plans, manages medical and disability aspects of worker's compensation injuries <dm>and</dm> is involved in claims adjustments for insurance companies. <dm>Also</dm> beginning trading today on the Big Board are El Paso Refinery Limited Partnership, El Paso, Texas, (ELP) and Franklin Multi-Income Trust, San Mateo, Calif., (FMI).

3. **Example 2:**

The Polish rat will eat well this winter. Tons of delectably rotting potatoes, barley and wheat will fill damp barns across the land <dm>as</dm> thousands of farmers turn the state's buyers away. Many a piglet won't be born as a result, <dm>and</dm> many a ham will never hang in a butcher shop. <dm>But</dm> with inflation raging, grain in the barn will still be a safer bet for the private farmer than money in the bank. Once again, the indomitable peasant holds Poland's future in his hands. <dm>Until</dm> his labor can produce a profit in this dying and distorted system, even Solidarity's sympathetic new government won't win him over.

Your Task:

- Read the following German text.

- Identify all DMs as per the guidelines above.

- Enclose each DM within <dm> tags.

- Ensure that the rest of the text remains unchanged.

German Text:

B Frequently confused relations

Relations	Ann. A	Ann. B
elaboration-evidence	13	10
cause-list	7	8
cause-reason	5	5
cause-evidence	6	4
cause-elaboration	5	4
conjunction-list	26	2
joint-list	2	7
concession-antithesis	8	0

Table 5: Frequent two-sided (top) and one-sided (bottom) relation confusions in Kobalt

Relations	Ann. A	Ann. B
elaboration-interpretation	15	11
elaboration-nonvol-cause	15	11
elaboration-circumstance	14	11
elaboration-nonvol-result	12	11
elaboration-background	6	12
elaboration-conjunction	11	6
circumstance-condition	10	5
elaboration-preparation	8	7
justify-motivation	7	5
joint-conjunction	22	1
nonvol-cause-nonvol-res	12	3
joint-list	11	1
summary-preparation	8	1
nonvol-cause-circumstance	7	1

Table 6: Frequent two-sided (top) and one-sided (bottom) relation confusions in NLDT