Where Frameworks (Dis)agree: A Study of Discourse Segmentation

Maciej Ogrodniczuk[1] **Anna Latusek**[1] **Karolina Saputa**[1,†] **Alina Wróblewska**[1]
Daniel Ziembicki[1] Bartosz Żuk[1] Martyna Lewandowska[1] Adam Okrasiński[2]
Paulina Rosalska[3] Anna Śliwicka[1] Aleksandra Tomaszewska[1] Sebastian Żurowski[3]
[1]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[2]Faculty of Psychology, University of Warsaw, Warsaw, Poland.
[3]Faculty of Humanities, Nicolaus Copernicus University, Toruń, Poland.
{m.ogrodniczuk, a.latusek, alina, d.ziembicki, b.zuk}@ipipan.waw.pl

Abstract

This study addresses the fundamental task of discourse unit detection – the critical initial step in discourse parsing. We analyze how various discourse frameworks conceptualize and structure discourse units, with a focus on their underlying taxonomies and theoretical assumptions. While approaches to discourse segmentation vary considerably, the extent to which these conceptual divergences influence practical implementations remains insufficiently studied. To address this gap, we investigate similarities and differences in segmentation across several English datasets, segmented and annotated according to distinct discourse frameworks, using a simple, rule-based heuristics. We evaluate the effectiveness of rules with respect to gold-standard segmentation, while also checking variability and cross-framework generalizability. Additionally, we conduct a manual comparison of a sample of rule-based segmentation outputs against benchmark segmentation, identifying points of convergence and divergence.

Our findings indicate that discourse frameworks align strongly at the level of segmentation: particular clauses consistently serve as the primary boundaries of discourse units. Discrepancies arise mainly in the treatment of other structures, such as adpositional phrases, appositions, interjections, and parenthesised text segments, which are inconsistently marked as separate discourse units across formalisms.

1 Introduction

Several linguistic discourse theories have been developed to model the structure and coherence of texts, such as Rhetorical Structure Theory (RST, Mann and Thompson, 1988; Taboada and Mann, 2006) and Penn Discourse Treebank (PDTB, Prasad et al., 2008). Each of them proposes a distinct set of discourse relations and discourse unit types, as well as specific assumptions about the hierarchical or relational structures that reflect discourse organization. These theoretical frameworks form the basis for a range of practical implementations, e.g., the creation of annotated datasets and the development of discourse parsers.

Computational approaches to discourse adopt various taxonomies, but they typically divide discourse parsing into two main subtasks: the identification of discourse units and the classification of the relations between them (Braud et al., 2023). Discourse units (DUs) correspond to spans of text that convey discourse-relevant content, such as events, states, facts, and propositions. Discourse relations (DRs), in turn, connect DUs and assign labels to them based on a predefined taxonomy, with categories such as contrast, elaboration, or purpose.

In this study, we focus on the initial and foundational stage of discourse parsing – DU detection. This step is particularly critical, as it significantly impacts the overall accuracy of the subsequent discourse parsing task – discourse relation classification. We review how various discourse frameworks conceptualize and define DUs, with particular attention to the taxonomies they propose (Section 2). We then verify how these theoretical definitions are operationalized in practice by evaluating their implementations in existing datasets. To support this analysis, we propose a simple, rule-based heuristics for discourse segmentation (Section 3). We apply it to several English discourse datasets to validate its effectiveness and generalizability (Section 4). Furthermore, we conduct a detailed manual analysis of segmentation differences between datasets representing different discourse formalisms, aiming to identify where these frameworks align and where they diverge (Section 5).

Our current objective is to compare various approaches to discourse segmentation in real English datasets and to identify their commonalities and distinctions. Ultimately, we aim to develop a dis-

[†]To the memory of Karolina, tragically deceased on August 14, 2025.

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 182–196
November 9, 2025 ©2025 Association for Computational Linguistics

course segmentation strategy that is both theoretically grounded and practically robust, i.e., consistent across various discourse frameworks and applicable to multiple languages.

Contributions

- A systematisation of discourse frameworks and their associated resources.
- An interpretable, syntax-based discourse segmentation approach.
- An analysis of segmentation inconsistencies in multiple English discourse datasets.

2 Discourse Segmentation in Various Formalisms

Based on Marcu (2000), Wolf and Gibson (2005) point out that there is no agreement on the notion of discourse unit: "discourse segments should be prosodic units (Hirschberg and Nakatani, 1996), others argue for intentional units (Grosz and Sidner, 1986), phrasal units (Longacre, 1983; Lascarides and Asher, 1993; Webber et al., 1999), or sentences (Hobbs, 1985)."

In general, existing discourse representation formalisms adopt one of two approaches to text segmentation: DUs are either independent of DRs, or they function only as relation arguments. The former implies that a text can contain DUs that are not part of any DR (but may be part of other coherence relations). In contrast, the latter approach limits the definition of DUs to those that participate in DRs, omitting segments that do not conform to a particular formalism.

Further questions to consider include whether DUs can overlap or form hierarchies within each formalism, whether they cover the text completely, and whether discourse markers are part of DUs. In this section, we will address each of these questions by reviewing the most prominent existing discourse representation formalisms and summarizing the findings of this review in Table 2 in Appendix A.

2.1 Non-implemented Discourse Formalisms

Hobbs' Theory of Discourse Coherence (HTDC) Hobbs (1985) proposes an early text coherence theory that treats the connection of discourse units through discourse relations as an indicator of text coherence. This theory is closely linked to coreference, meaning subsequent sentences in a coherent text should refer to the same entity (Hobbs, 1979). Clauses consist of predicates about the entities referred to in the text.

According to Hobbs's theory, a DU is a sentential unit, which is also known as *a segment of discourse*. A segment of discourse is a set of clauses and other sentential units. Every clause is a sentential unit. Two segments form a segment of discourse if they are connected by a relation and assertions of predicates in both segments can be connected into one set of assertions. Hobbs's theory is an attempt at a computable and implementable theory of discourse. However, the idea of assertions – the propositions in clauses that are asserted in constructing larger sentential units – or even the possibility to parse a clause into a set of assertions is difficult to implement (Hobbs, 1985, 2013).

Cognitive Approach to Coherence Relations (CCR) CCR, as elaborated by Sanders et al. (1992, 1993), represents a functional grounded framework for understanding discourse coherence. CCR emphasizes the cognitive processes and constraints that underlie how language users identify, categorize, and interpret coherence relations between discourse units.

Key aspects of the situation in CCR include the description of reality: objective relations in CCR typically pertain to DUs that describe situations or events that occur in the real world or in the world described by the text. Clues for the segmentation procedure are that DUs must be small enough to be a single information unit and interpretable on their own. In objective relations, the speaker, or author, merely reports facts and is not actively involved in constructing or evaluating the relation itself. This is evident, for example, in causal relations based on real-world causality, where one situation or event is presented as the cause of another. The concept of situation in CCR thus helps distinguish objective relations from subjective relations, which typically express the speaker's opinion, argument, or evaluative stance. Historically, one may have referred to these as "semantic" and "pragmatic" relations, respectively.

ISO 24617:2: Dialogue Acts The ISO standard for Dialogue Acts (Bunt et al., 2012) provides a framework for semantic annotation of dialogues. It differs significantly from other discourse segmentation approaches by adopting a fundamentally functional perspective on discourse units.

The primary segmentation units are dialogue acts, interpreted in terms of their communicative functions and semantic content, and distributed across multiple dialogue dimensions (e.g., Task Management, Feedback, Turn Management).

2.2 Discourse Frameworks Implemented as Datasets

GraphBank Wolf and Gibson (2005) present an annotation method for discourse coherence and evidence that trees are not an appropriate form for representing discourse structure. The authors argue that coherence structure can be represented in the form of a graph, in which nodes represent *discourse segments* and edges represent the discursive relations connecting these segments.

Discourse segments (aka DUs) correspond to clause units, non-restrictive relative clauses, and modifying prepositional phrases. The authors exclude complex nouns or verb phrases and restrictive relative clauses as DUs. Discourse units are typically marked by coordinating and subordinating conjunctions and punctuation marks. However, the conjunction 'and' does not mark the boundaries of segments if it connects nominal expressions and verb groups. As separate DUs, the authors also distinguish attributions that enable the distinction between different sources that comment on the same event. Attributions may be separated only if the attributed material is a complementizer phrase.

Prague Discourse Treebank 4.0 (PDiT) Intra- or inter-sentential discourse relations in PDiT 4.0 (Synková et al., 2024), labeled with semantic-pragmatic types, are determined by *explicitly* expressed discourse connectives that link exactly two *discourse arguments* (aka DUs). The primary and secondary discourse connectives are distinguished.

Since PDiT is built upon the Prague Dependency Treebank (Hajič et al., 2020), each DU is anchored in a single node of a tectogrammatical (deep-syntactic) tree representing a sentence, typically the root of the corresponding subtree. As a result, DUs correspond to text spans centered around a finite verb (i.e., the root node), with their boundaries determined by the extent of the subtree.

ISO 24617:8 ISO 24617-8 (International Organization for Standardization, 2016) is a standard created in 2016 for annotation of local discourse relations in any language or genre.

The basic discourse unit is *situation*, which covers any eventuality, fact, proposition, condition, belief, or dialogue act that can be realized by a clause, nominalization, full sentence, utterance, or extended DU. The standard is deliberately neutral on span adjacency: an argument may be minimal or extended, continuous or discontinuous, provided it denotes the intended situation. Relations may be symmetric, assigning identical roles to both arguments, or asymmetric, assigning distinct roles. Crucially, relations are defined independently of the presence or absence of discourse markers.

2.3 Discourse Frameworks Implemented as Tools

Rhetorical Structure Theory (RST) Discourse units are essentially sentences, coherent fragments of text characterized by functional integrity (Mann and Thompson, 1988, 248–249). The minimum unit refers to spans: nuclei and their satellites. The nuclei are crucial for maintaining the coherence of the text. Determining the relationship between nucleus units and satellite units allows the analyst to present schemata, which can then be used to represent RST structures in trees. To define the smallest DU, later works referring to RST began to use the term *elementary discourse unit* (EDU, Carlson et al., 2001).

RST has proven to be very useful for both linguistics and natural language processing, as numerous RST parsers have been developed for various languages, e.g. Carlson et al. (2001), Hernault et al. (2010), Cardoso et al. (2011), Stede and Neumann (2014), Iruskieta and Zapirain (2015). It should be noted that the assumptions of RST presented by Mann and Thompson (1988), including those concerning the segmentation of text, have been modified, sometimes significantly.

Penn Discourse Treebank (PDTB) PDTB (Prasad et al., 2008) does not explicitly formulate a definition of a DU. Instead, the framework is grounded in a lexically driven, minimal-pair approach, in which the primary focus is the discourse relation. This relation annotated between two spans of text, Arg1 and Arg2, is linked by a discourse connective (e.g. *because*, *and*, *since*), either explicitly present or implicitly inferred from the context. In the annotation process, the connective is identified first, and then the relation built around it is labeled.

The spans are usually equivalent to clauses; however, the overarching principle in PDTB is that they must be interpretable within the context of a discourse relation. As a result, they may extend beyond a single sentence or consist of only part of a clause. Since PDTB focuses on annotating discourse connectives rather than predefined discourse units, it employs a function-based, non-overlapping segmentation with no hierarchy or nesting.

Segmented Discourse Representation Theory (SDRT) SDRT (Lascarides and Asher, 1993; Asher and Lascarides, 2005) is a framework for modeling discourse semantics that extends Discourse Representation Theory.

The most basic building blocks are called *Elementary Discourse Units* (EDUs). An EDU is defined as the smallest unit of text or dialogue that is semantically independent enough to participate in discourse relations. Typically, this corresponds to a simple clause (or subclause), a complete utterance in dialogue, or sometimes a larger phrase when it functions as a standalone informational unit. Sometimes EDUs are combined in complex discourse units (CDUs).

3 Syntax-based Discourse Segmentation Across Frameworks

3.1 Preliminaries and Rationale

A wide range of discourse frameworks exists (see Section 2), each differing significantly in its theoretical foundations and descriptive conventions. However, it remains an open question to what extent their practical implementations – specifically, annotated datasets and discourse parsers – also diverge. Do these resources reflect fundamental conceptual differences, or do they share underlying similarities? In this study, we aim to shed light on this issue by focusing on the task of discourse segmentation, which represents the initial step in discourse parsing. Using a simple, rule-based heuristics (see Section 3.2), we perform segmentation on a selection of English datasets annotated according to various discourse frameworks.

Predicative-argument structures – comprising sentence-level predicates and their associated arguments – form the backbone of meaning representation in natural language. These structures not only organize the semantics of individual sentences and clauses, but also serve as building blocks for larger discourse-level constructions. Due to their universality across languages, predicate-argument structures provide a promising foundation for identifying DUs.

We assume that boundaries of DUs align with the surface realization of predicate-argument structures. Without a doubt, these surface realizations vary significantly across languages, influenced by factors such as word order, morphology, and accepted argument or predicate ellipses. Despite this variation, Universal Dependencies (UD, de Marneffe et al., 2021) approximate predicate-argument relations using a cross-linguistically consistent schema. This makes UD trees a practical and theoretically grounded resource for implementing a discourse segmentation method that generalizes across typologically diverse languages.

Universal Dependencies are structured around two fundamental linguistic concepts: the nominal, typically used to represent entities, and the clause, generally used to denote events and states. Clauses consist of a main predicate along with its arguments and modifiers. They may function as either independent sentences or embedded clauses (i.e. realizations of arguments or modifiers). Most of the formalisms discussed in Section 2 define discourse units roughly as clauses. Identifying clauses within UD trees makes it possible to extract corresponding DUs and thus to segment discourse.

3.2 UD-based Discourse Segmentation Rules

Building on the above-mentioned observations, and in line with prior research (Braud et al., 2017; Desai et al., 2020), we assume that discourse relations primarily hold between DUs realized as clauses. Consequently, we define a set of simple UD-based rules to identify clause structures. This clause identification serves as the foundation for detecting DU boundaries in a consistent and linguistically motivated way.

R1. Clauses UD distinguishes several dependency types that are realized as clausal structures and can be directly used for discourse segmentation. In particular, the heads of clauses are typically marked with the following types:

- *root* – the root of a sentence,
- *ccomp* – a clausal complement of a verb or adjective; a finite clause with an internal subject,
- *advcl* – an adverbial clause modifying a predicate or modifier word,
- *acl* – an adnominal clause, i.e., a finite or non-finite clause modifying a nominal (nominal postmodifier).

The full subtree headed by each of these dependency types (see Ex. (1) and (2)) corresponds to a distinct DU.

(1) ⟨You're so stupid$_{root}$⟩ ⟨thinking$_{advcl}$⟩ ⟨I spent$_{ccomp}$ the night.⟩

(2) ⟨This is a trend$_{root}$⟩ ⟨that bears$_{acl:relacl}$ more scrutiny⟩ ⟨than it has received$_{acl}$.⟩

R2. Parataxis The *parataxis* relation captures constructions in which clauses or constituents are placed side by side without an explicit coordination or subordination structure. Each subtree governed by a *parataxis* head is treated as a distinct separate segment (Ex. (3)).

(3) ⟨As each task becomes$_{adcvl}$ more specialized,⟩ ⟨Smith noted$_{parataxis}$,⟩ ⟨it engages$_{root}$ less of the person.⟩

R3. Relative clauses The dependency type *acl:relcl* is used to annotate relative clauses (Ex. (4), (5)), as well as subordinate clauses introduced by a relative pronoun that simultaneously serves as an argument of the main predicate (Ex. (6)). All such constructions are treated as individual DUs.

(4) ⟨Such a scenario may be found$_{root}$ in different situations,⟩ ⟨including when one studies$_{acl:relcl}$ a language in a classroom...⟩

(5) ⟨I'm supposed$_{root}$ to trust you every time⟩ ⟨you tell$_{acl:relcl}$ the truth.⟩

(6) ⟨But how am I supposed$_{root}$ to know⟩ ⟨when you're telling$_{acl:relcl}$ the truth?⟩

R4. Coordination Clauses that are coordinated with other clauses, such as those mentioned above, along with *parataxis* or *xcomp* (a clausal complement of a verb or adjective with an obligatorily controlled subject), are always treated as separate DUs. Similarly, clauses with elided predicates are also considered independent segments (Ex. (7)).

(7) ⟨Finally, in some cases a gain in performance has been observed$_{root}$: after 1.5 years of limited exposure in one study [...],⟩ ⟨and [observed]$_{ellipsis}$ in another study after 2 years,⟩ ⟨though [are observed]$_{ellipsis}$ only for some abilities [...]⟩

R5. Parenthesis Parenthesized content, e.g., bibliographic references, is treated as a separate DU. In contrast, bracketed elements that function as appositions, predicate arguments or modifiers are not segmented (Ex. (8)). This distinction reflects the assumption that bracket usage in such cases carries a higher-level, likely pragmatic, interpretation rather than signaling a distinct DU.

(8) ⟨Higher levels of proficiency (or exposure) may be associated$_{root}$ with less attrition⟩ ⟨[17$_{dep}$], [18], [21], [23]⟩ ⟨or even with no observed losses$_{discontinuation}$ ⟨[21$_{dep}$].⟩

While other punctuation marks (i.e., periods, semicolons, and commas) can align with DU boundaries, they are not dependable cues for segmentation. They may be inconsistent or redundant, or reflect higher-level phenomena, e.g., at the pragmatic or stylistic level, rather than indicating clear DUs boundaries.

Discourse connectives, such as complementizers and subordinating conjunctions (annotated as *mark*), often introduce embedded DUs. Similarly, discourse markers (*discourse*) may start new DUs. However, at this preliminary stage, these elements are not segmented separately. Our goal is to examine how they are handled across different discourse frameworks to identify a consistent and unified approach to their treatment.

3.3 Discourse Segmentation Algorithm

A key component of the discourse segmentation algorithm involves identifying which tokens within a sentence correspond to DU heads, i.e., tokens that anchor individual DUs. The selection of head tokens is guided by:

- their dependency relation types,
- their part-of-speech tags,
- and the dependency types and part-of-speech tags of their particular dependent tokens.

Once the head tokens are identified (e.g., *likes* and *lost* in Figure 1), the next step involves determining the token span of each DU using the structure of the UD subtree headed by the identified tokens. A segment X consists of its head token x and possibly all tokens contained within the subtree rooted at x, denoted as T_x. However, if another head token y, representing a separate DU, is nested within T_x, the span of segment X is restricted to tokens in T_x that are not part of the subtree rooted at y (i.e., T_y). In cases where segment X continues after the interruption caused by segment Y, a *discontinuity* discourse relation can be directly added, based on the hierarchical structure of T_x.

We adopt a rule-based approach to discourse segmentation rather than training a dedicated model or relying on large language models (LLMs). This decision is motivated by several factors. First, our objective is to develop a unified segmentation method applicable across multiple discourse datasets, annotated according to different discourse frameworks. Training a model on a single dataset would likely result in overfitting to the specific annotation conventions and discourse structure assumptions of that dataset. On the other hand, training a single model on a compilation of all available datasets introduces the risk of learning from inconsistent or conflicting instances, which may lead to poor or unpredictable performance. Furthermore, discourse segmentation models show limited generalizability, even across datasets in the same language and framework (Muller et al., 2019). Moreover, while LLMs can be fine-tuned or prompted to perform discourse segmentation, their output is often difficult to interpret without a complete manual revision or further postprocessing. In particular, it is difficult to systematically verify which syntactic or semantic cues underlie the identified DUs. Additionally, LLMs are prone to unintended alterations in the input, making them unreliable for the current study, where preserving the original text is crucial. Given these constraints, a rule-based approach offers a controlled and interpretable framework that facilitates cross-formalism comparison.

4 Evaluation

As a basic proof of concept, we evaluate our approach on English datasets from DISRPT 2023 (Braud et al., 2023) Task 1 (Treebanked Segmentation). The datasets span three discourse frameworks (RST, SDRT, and DEP) across diverse domains:

- `eng.dep.scidtb` (Yang and Li, 2018) contains scientific abstracts,
- `eng.dep.covdtb` (Nishida and Matsumoto, 2022) contains COVID-19 research abstracts,
- `eng.rst.rstdt` (Lynn Carlson, 2002) contains Wall Street Journal articles from the Penn Treebank,
- `eng.sdrt.stac` (Asher et al., 2016) contains chat dialogues from the Settlers of Catan game,
- `eng.rst.gum` (Zeldes, 2017) contains mixed genres including essays, interviews, and online forum discussions.

It is important to articulate that the DEP datasets follow the RST annotation guidelines for DU segmentation, and their discourse relation set is based on PDBT.

Using the provided tokenization and dependency trees from the DISRPT repository[1], we apply our rule-based discourse segmenter to obtain discourse segmentation and report precision, recall, and F1 scores in Table 1.

We compare against two DISRPT 2023 participants: DisCut (Metheniti et al., 2023) and HITS (Liu et al., 2023). Both systems used pretrained language models (`XLM-RoBERTa` for DisCut, `RoBERTa-large` for HITS) fine-tuned separately on each dataset. For the out-of-distribution `eng.dep.covdtb` dataset, which only provided dev and test splits, both teams used models trained on `eng.dep.scidtb`.

While our rule-based approach trails DisCut and HITS by 5.61 and 5.17 F1 points respectively, it demonstrates strong generalization ability. It achieves consistent performance across three different discourse frameworks and vastly different domains. The performance gap narrows significantly on out-of-distribution data: on `eng.dep.covdtb`, our system comes within just 2.01 points of HITS and 3.84 of DisCut. This convergence highlights a key advantage – our single rule-based system comes close to state-of-the-art performance without requiring data labeling and training separate model for each domain and discourse framework.

These results may suggest that the three discourse frameworks do not differ much concerning segmentation, indicating that they share common assumptions about how discourse should be divided into segments. Furthermore, the observed alignment between our rule-based segments and those in the datasets supports our intuition that DUs generally correspond to clauses. However, despite this apparent similarity, some differences remain between our rule-based segmentation heuristics and the datasets, and they should be examined more

[1] `https://github.com/disrpt/sharedtask2023`

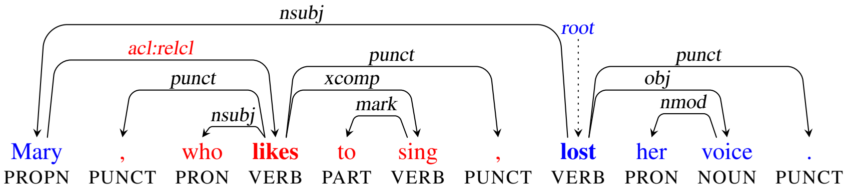

Figure 1: A dependency tree with two head tokens and the corresponding DUs highlighted in blue and red.

Dataset	Our			DisCut			HITS		
	P	R	F1	P	R	F1	P	R	F1
`eng.rst.rstdt`	87.56	87.00	87.28	97.21	98.04	**97.62**	96.46	97.66	97.06
`eng.rst.gum`	90.76	93.14	91.93	94.59	96.42	**95.50**	95.08	95.29	95.19
`eng.sdrt.stac`	87.92	90.37	89.13	95.75	94.70	95.22	96.71	95.09	**95.89**
`eng.dep.scidtb`	90.84	90.88	90.86	94.96	95.18	**95.07**	94.77	95.09	94.93
`eng.dep.covdtb`*	89.60	87.02	88.29	94.04	90.31	**92.13**	90.22	90.38	90.30
Mean	89.34	90.20	89.50	95.51	94.93	**95.11**	94.65	94.70	94.67

Table 1: Segmentation precision, recall and F1 score on English DISRPT 2023 datasets (Braud et al., 2023) comparing our approach to DisCut (Metheniti et al., 2023) and HITS (Liu et al., 2023) (Treebanked track). *indicates out-of-distribution datasets without training data.

closely. It is important to ask whether certain segments result from formalism-specific segmentation principles and are unique to a particular theoretical framework. Addressing these questions is essential for understanding the concept of DUs and the underlying theoretical assumptions in each discourse framework.

5 Analysis of Segmentation Discrepancies

After applying our heuristics to five different datasets, we conduct a comparative manual analysis of their samples (i.e. about 10 sentence pairs with segmentation discrepancies from each dataset). Its main goal is to identify the areas where rule-based segmentation and gold-standard segmentation diverge, assess the extent of the differences between them, and determine whether these discrepancies significantly exceed the boundaries of the DU definition we adopted.

We argued that the segmentation method – using clause boundaries as the foundation for defining DUs – is grounded in solid linguistic principles. Our analyses largely confirm this assumption, as the majority of approaches converge with our clause-level segmentation. Differences occur rather in individual cases than globally and are caused by various factors, often due to incorrect morphosyntactic annotation.

Concerning discrepancies, they are grouped into true discrepancies, reflecting systematic differences in segmentation principles (see Section 5.1), and other discrepancies, caused by preprocessing errors and gold data inconsistencies (see Section 5.2).

5.1 True Discrepancies

R1. Adverbial clauses (*advcl*) The highest level of agreement concerns adverbial clauses modifying a predicate or modifier word, which in most datasets – similarly to our approach – are segmented. The only exception is `eng.sdrt.stac`, where adverbial clauses are very consistently not segmented (see B.1).

R1. & R3. Adnominal clauses (*acl*) and relative clauses (*acl:relcl*) A similar degree of overlap (high agreement) between our segmentation and the compared datasets can be observed in the case of all adnominal clauses, including relative clauses. However, `eng.sdrt.stac` again stands out, where such segmentation is regularly not performed (see B.2).

R1. Clausal complements (*ccomp*) The segmentation of complement clauses brings similar conclusions to the previous ones. In situations involving a verb complement, two UDs are distinguished in most datasets. The only exception shows the `eng.sdrt.stac` dataset (see B.3).

R4. Verb coordination In the case of verb coordination, there is no longer such agreement. In a larger number of datasets, when there is a single subject and a series of coordinated verbs related to it, no division into separate parts may be made. This is particularly noticeable in `eng.dep.scidtb`). Our approach always segments each of the predicates into separate DUs, no matter whether they are coordinated (see B.4).

R5. Apposition in brackets This is the area of the least agreement between our approach and the others. In RST-type datasets, such as `eng.rst.rstdt` and `eng.rst.gum`, all information in brackets, including appositions, is frequently annotated as separate DUs (see B.5, B.6). In DEP- and SDRT-type approaches, similarly to our heuristics, appositions in the brackets are not marked as separate DU.

Punctuation As we mentioned in Section 3, we do not rely on punctuation in our segmentation. However, in some datasets, colons or semicolons were used to mark DU boundaries. This is particularly evident in RST-type datasets, where a colon after an introductory phrase always signals a new unit (see B.7).

Interjections Interjections, such as 'right', 'well', 'no', 'sorry', 'hi', are marked as separate DUs in `eng.sdrt.stac`. In contrast, both `eng.erst.gum` and our approach treat them as part of an adjacent DU (see B.8).

Adpositional phrases In `eng.dep.covdtb`, gerunds (e.g. 'including', 'regarding') and gerunds followed by an adposition (e.g. 'according to') are annotated as adpositional phrases that constitute separate DUs (see B.9).

5.2 Other Discrepancies

Discourse datasets are annotated with morphosyntactic information, including sentence and token segmentation, part-of-speech tags, morphological features, and dependency trees. These annotations were either derived from existing resources or predicted automatically. Since automatic preprocessing is prone to errors, it can negatively impact the quality of discourse segmentation.

Preprocessing errors Some of the observed mismatches can be traced to inaccuracies in POS tagging or dependency parsing errors in the gold-standard data. For instance, multiple errors occur in `eng.dep.covdtb`, including misidentification of sentence predicates (see B.10), incorrect analyses of coordination structures, erroneous assignment of part-of-speech tags, etc.

Inconsistencies in gold standard In RST-type approaches, appositions are treated as separate DUs only when enclosed in parentheses. When they are set off by commas, they remain part of the main clause. This inconsistent treatment of appositions leads to differences in segmentation between this dataset and our approach (described in the previous section).

Over-segmentation Our rules sometimes produce unnecessary splits, which becomes particularly evident when it comes to identifying and segmenting spans containing proper names, titles, or compound nouns (see B.11).

6 Conclusions

All of our analyses – initial, focusing on the most popular approaches to discourse segmentation; subsequent, examining how these theories are implemented in annotated datasets; and concluding, comparing our proposal of the UD-based discourse segmentation with already existing ones – have allowed us to preliminarily confirm the assumption that, in practice, all these approaches share a significant number of common characteristics.

Moreover, our UD-based approach to discourse segmentation, grounded in simple and clear rules, has yielded very promising results. We are aware that our rule-based approach does not reach the performance of state-of-the-art methods. However, surpassing these methods was not our primary objective. It is nevertheless worth emphasizing that segmentation based on five simple rules, applied uniformly across all tested datasets annotated according to different formalisms, approaches the performance of models trained separately for each dataset.

Our main goal was to investigate whether syntax can contribute to discourse segmentation, and the results suggest that it constitutes a key factor in identifying discourse units. This finding stands in partial contrast to the results reported by Braud et al. (2017). However, a direct comparison is difficult, given the substantially different assumptions and experimental setups. This issue, therefore, calls for further dedicated research.

Our findings encourage us to research the topic further and set a goal of developing a universal, cross-approach method for detecting DUs. Naturally, certain areas remain that warrant particular attention. They will need to be addressed in greater depth in the next stages of our study, e.g., manual comparative analyses, in which we focused on identifying as many points of divergence in segmentation as possible, revealed discrepancies primarily at the level of segmenting certain types of subordinate clauses, parataxis, appositions or units' discontinuity. These are areas that require special attention in future work.

It should also be noted that our comparative analysis of the proposed UD-based discourse segmentation was conducted across a selection of datasets that, in our view, appeared to be the most representative. The choice of these particular datasets was motivated by our understanding of discourse as a phenomenon in which the text operates as a whole. In this perspective, discourse refers to a specific way of organizing the text holistically, where every element plays a defined role and, therefore, cannot be omitted in segmentation.

Nonetheless, we acknowledge that developing a universal approach to discourse segmentation will require further evaluations and analyses of cross-linguistic and cross-dataset similarities and differences in future work.

Key Insights

- High level of agreement on DUs segmentation across the evaluated datasets.
- Promising results of the UD-based method for DUs segmentation, demonstrating strong generalization ability.
- Potential for unifying rules for cross-linguistic discourse segmentation.

Acknowledgments

This research was funded in whole by the National Science Centre, Poland, grant 2023/50/A/HS2/00559 ("Universal Discourse: a multilingual model of discourse relations").

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2005. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Does syntax help discourse segmentation? Not so much. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2432–2442, Copenhagen, Denmark. Association for 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.

H Bunt, J Alexandersson, J-W CHOE, Alex Chengyu FANG, K Hasida, V Petukhova, A Popescu-Belis, and D Traum. 2012. Iso 24617-2: 2012: Language resource management–semantic annotation framework (semaf)–part 2: Dialogue acts.

Paula C. F Cardoso, Erick G Maziero, Maria Lucía R. Castro Jorge, Eloize M. R Seno, Ariani Di Felippo, Lucia H. M Rino, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. Cstnews: a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Brazilian Symposium in Information and Human Language Technology - STIL*. SBC.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, page 1–10, USA. Association for Computational Linguistics.

Ta-Chung Chi and Alexander Rudnicky. 2022. Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–335, Edinburgh, UK. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Takshak Desai, Parag Pravin Dakle, and Dan Moldovan. 2020. Joint learning of syntactic features helps discourse segmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1073–1080, Marseille, France. European Language Resources Association.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

Grigorii Guz, Patrick Huber, and Giuseppe Carenini. 2020. Unleashing the power of neural discourse parsers - a context and structure aware approach using large scale pretraining. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3794–3805, Barcelona, Spain. International Committee on Computational Linguistics.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Hugo Hernault, Helmut Prendinger, David duVerle, and Mitsuru Ishizuka. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue & Discourse*, 1(3).

Julia Hirschberg and Christine H. Nakatani. 1996. A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Santa Cruz, California, USA. Association for Computational Linguistics.

Jerry R. Hobbs. 1979. Coherence and Coreference*. *Cognitive Science*, 3(1):67–90.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information (CSLI), Stanford, CA.

Jerry R. Hobbs. 2013. ACL lifetime achievement award: Influences and inferences. *Computational Linguistics*, 39(4):781–798.

International Organization for Standardization. 2016. ISO 24617-8:2016 Language resource management — Semantic annotation framework (SemAF) — Part 8: Semantic relations in discourse (DR-Core). Standard ISO 24617-8:2016, International Organization for Standardization, Geneva, Switzerland.

Mikel Iruskieta and Beñat Zapirain. 2015. Euseduseg: A dependency-based edu segmentation for basque. *Proces. del Leng. Natural*, 55:41–48.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. *Preprint*, arXiv:1011.0835.

Wei Liu, Yi Fan, and Michael Strube. 2023. HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Robert E. Longacre. 1983. *The Grammar of Discourse*. Plenum, Mew York.

Mary Ellen Okurowski Lynn Carlson, Daniel Marcu. 2002. RST Discourse Treebank LDC2002T07.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text & Talk*, 8:243–281.

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

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. 2024. Polish Discourse Corpus (PDC): Corpus design, ISO-compliant annotation, data highlights, and parser development.

In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12829–12835, Torino, Italy. ELRA and ICCL.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Lynn Webber. 2008. The Penn Discourse TreeBank 2.0. In *International Conference on Language Resources and Evaluation*.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, 4(2):93–134.

Purificaç ao Silvano, Jo ao Cordeiro, Ant'onio Leal, and Sebasti ao Pais. 2023. DRIPPS: a corpus with discourse relations in perfect participial sentences. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 470–481, Vienna, Austria. NOVA CLUNL, Portugal.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *International Conference on Language Resources and Evaluation*.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the Prague Discourse Treebank 3.0. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1270–1279, Torino, Italia. ELRA and ICCL.

Pavlína Synková, Jiří Mírovský, Marie Paclíková, Lucie Poláková, Magdaléna Rysová, Veronika Scheller, Jana Zdeňková, Šárka Zikánová, and Eva Hajičová. 2024. Prague Discourse Treebank 4.0 (PDiT 4.0).

Maite Taboada and William C. Mann. 2006. Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. Llamipa: An incremental discourse parser. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6418–6430, Miami, Florida, USA. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2015. A Refined End-to-End Discourse Parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning – Shared Task*, pages 17–24, Beijing, China. Association for Computational Linguistics.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse Relations: A Structural and Presuppositional Account Using Lexicalised TAG. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, College Park, Maryland, USA. Association for Computational Linguistics.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. 2005. Discourse Graphbank. LDC Catalog No.: LDC2005T08.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM Corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Properties of Discourse Units Across Major Discourse Frameworks

Discourse Approach	DU Name	DU Form	FTC	OV	SC	I	Dataset / Parser
Hobbs' Theory of Discourse Coherence (HTDC)	segment / sentential unit	clause / set of clauses	N	N	N	N	N
Cognitive Approach to Coherence Relations (CCR)	discourse segment	any text span, cognitively motivated	N	Y	Y	N	N
ISO 24617-2 (Dialogue acts)	functional segment	Turns, utterances, sub-utterances	Y	Y	N	Y	N
GraphBank	discourse segment	clause	Y	N	Y	Y	Y • Discourse Graphbank (Wolf et al., 2005)
Prague Discourse Treebank 4.0 (PDiT)	discourse argument	Root nodes of tectogrammatical subtrees (i.e. heads of finite clauses; text spans centred around finite verbs)	N	N	Y	Y	Y • PDiT 4.0 (Synková et al., 2024)
ISO 24617-8	situation	eventuality, fact, proposition, condition, belief or dialogue act	N	Y	Y	Y	Y • DRIPPS (Silvano et al., 2023) • PDC (Ogrodniczuk et al., 2024)

Discourse Approach	**DU Name**	**DU Form**	**FTC**	**OV**	**SC**	**I**	**Parser / Dataset**
Rhetorical Structure Theory (RST)	elementary discourse units (EDU)	nucleus or sattelite span (essentially clauses)	Y	N	N	Y	Y • DPLP parser (Ji and Eisenstein, 2014), • Hilda (Hernault et al., 2010), • RST parser (Guz et al., 2020), • Top-Down parser (Koto et al., 2021), • RST parser (Yu et al., 2022)
Penn Discourse Treebank (PDTB)	argument	a minimal span of text that conveys a single discourse function (most often clauses)	N	N	Y	Y	Y • PDTB parser (Lin et al., 2010), • CoNLL-2015 shared task parsers (Wang and Lan, 2015)
Segmented Discourse Representation Theory (SDRT)	discourse unit (EDU/CDU)	propositions / clauses	N	Y	Y	Y	Y • DDP parser (Liu and Chen, 2021), • SDD parser (Chi and Rudnicky, 2022), • Llamipa (Thompson et al., 2024)

Table 2: Properties of discourse units across major discourse representation frameworks. Abbreviations: Y – Yes, N – No, FTC – Full Text Coverage; OV – DU Overlap; SC – Separate Connective (i.e. connectives treated as separate units); I – Implementability (i.e., the feasibility of implementing the framework in a dataset or a discourse parser.

B Types of Segmentation Discrepancies with Corresponding Examples

ID	Category	Dataset	Gold segmentation	Our segmentation
B.1	Adverbial clauses	`eng.sdrt.stac`	⟨Please only start the game when all four participants are there⟩	⟨Please only start the game⟩ ⟨when all four participants are there⟩
B.2	Adnominal clauses, relative clauses	`eng.sdrt.stac`	⟨I respect popular music from the time in which it was actually musical.⟩	⟨I respect popular music from the time⟩ ⟨in which it was actually musical.⟩
B.3	Relative clauses	`eng.sdrt.stac`	⟨I didn't know you could have them.⟩	⟨I didn't know⟩ ⟨you could have them.⟩
B.4	Verb coordination	`eng.dep.scidtb`	⟨We observe, identify, and detect naturally occurring signals of interestingness in click transitions on the Web between source and target documents,⟩ ⟨which we collect from commercial Web browser logs.⟩	⟨We observe,⟩ ⟨identify,⟩ ⟨and detect naturally occurring signals of interestingness in click transitions on the Web between source and target documents,⟩ ⟨which we collect from commercial Web browser logs.⟩
B.5	Parenthesis	`eng.rst.gum`	⟨Also 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) **.**⟩	⟨Also 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) **.**⟩
B.6	Appositions in brackets	`eng.dep.scidtb`	(...) ⟨health practices⟩ ⟨(exercise, tobacco and alcohol consumption, sleep efficiency)⟩ ⟨and genetics contribute to CLI risk.⟩	(...) ⟨health practices (exercise, tobacco and alcohol consumption, sleep efficiency) and genetics contribute to CLI risk.⟩
B.7	Punctuation	`eng.rst.gum`	⟨Respondents were asked to indicate their race from among the following categories:⟩ ⟨White; Black or African American; Hispanic; American Indian or Native American; and Asian or Pacific Islander.⟩	⟨Respondents were asked to indicate their race from among the following categories: White; Black or African American; Hispanic; American Indian or Native American; and Asian or Pacific Islander.⟩
B.8	Interjections	`eng.sdrt.stac`	⟨we 're waiting for 2 other players⟩ ⟨right⟩ ⟨no⟩ ⟨sorry⟩	⟨we 're waiting for 2 other players right⟩ ⟨no sorry⟩

B.9	Adpositional phrases	`eng.dep.covid`	⟨Severe infections can lead to a variety of diseases,⟩ ⟨including poliomyelitis, aseptic meningitis, myocarditis and neonatal sepsis.⟩	⟨Severe infections can lead to a variety of diseases, including poliomyelitis, aseptic meningitis, myocarditis and neonatal sepsis.⟩
B.10	Preprocessing errors	`eng.dep.covid`	Phylogenetic analysis based on S3 gene showed that the Brazilian TReoV isolates$_{ccomp}$ clustered in a single group with 98-100% similarity to TReoV strains circulating in the United States.	
B.11	Compounds	`eng.rst.gum`	⟨Among his smaller works, the seventh Humoresque and the song "Songs My Mother Taught Me" are also widely performed and recorded.⟩	⟨Among his smaller works, the seventh Humoresque and the song "Songs⟩ ⟨My Mother Taught Me"⟩ ⟨are also widely performed⟩ ⟨and recorded.⟩

Table 3: Segmentation discrepancy types categorized by type of linguistic phenomenon.