

## Discourse Relations and Connectives in Higher Text Structure

**Lucie Poláková**

*Institute of Formal and Applied Linguistics  
Faculty of Mathematics and Physics, Charles University*

POLAKOVA@UFAL.MFF.CUNI.CZ

**Jiří Mírovský**

*Institute of Formal and Applied Linguistics  
Faculty of Mathematics and Physics, Charles University*

MIROVSKY@UFAL.MFF.CUNI.CZ

**Šárka Zikánová**

*Institute of Formal and Applied Linguistics  
Faculty of Mathematics and Physics, Charles University*

ZIKANOVA@UFAL.MFF.CUNI.CZ

**Eva Hajičová**

*Institute of Formal and Applied Linguistics  
Faculty of Mathematics and Physics, Charles University*

HAJICOVA@UFAL.MFF.CUNI.CZ

**Editor:** Vera Demberg

Submitted 12/2020; Accepted 06/2021; Published online 07/2021

### Abstract

The present article investigates possibilities and limits of local (shallow) analysis of discourse coherence with respect to the phenomena of global coherence and higher composition of texts. We study corpora annotated with local discourse relations in Czech and partly in English to try and find clues in the local annotation indicating a higher discourse structure. First, we classify patterns of subsequent or overlapping pairs of local relations, and hierarchies formed by nested local relations. Special attention is then given to relations crossing paragraph boundaries and their semantic types, and to paragraph-initial discourse connectives. In the third part, we examine situations in which annotators incline to marking a large argument (larger than one sentence) of a discourse relation even with a minimality principle annotation rule in place. Our analyses bring (i) new linguistic insights regarding coherence signals in local and higher contexts, e.g. detection and description of hierarchies of local discourse relations up to 5 levels in Czech and English, description of distribution differences in semantic types in cross-paragraph and other settings, identification of Czech connectives only typical for higher structures, or the detection of prevalence of large left-sided arguments in locally annotated data; (ii) as another type of contribution, some new reflections on methodologies of the approaches under scrutiny.

**Keywords:** local and global discourse coherence, Prague Discourse Treebank 2.0, discourse connectives, hierarchies, paragraphs

### 1. Introduction

In coherence-oriented discourse studies, the recognition of the distinction between local and global coherence dates back to the 1980's and one of the most compelling ways of its explanation is that “the main purpose of global coherence relations is to help eliminate locally coherent nonsense texts” (Unger, 2006, Samet and Schank, 1984). According to them, global coherence is the connectivity

between the main events of the text (scripts, plans and goals) and the global relations hold independently of the local coherence relations between discourse segments. In automatic discourse processing, the local and the global coherence models are also known as shallow and deep discourse analyses/parsing, respectively (Prasad et al., 2010). Both types of approaches deal with determination and description of semantic and pragmatic relations between individual text units but they differ in the focus of the analyses and the methodology.

Local coherence models proceed bottom-up, from the smallest discourse units within a single sentence (clauses, even nominalized events, states etc.) and across the sentence boundary. Emphasis is put on the description of semantico-pragmatic relations between every two consecutive units and on the identification of lexical cues anchoring these relations (mostly discourse connectives; where there is no such surface cue, the relation is called implicit). The coherence analysis in this way describes how every discourse unit is related to the previous one, there are no hierarchies postulated and no claims about the shape of the overall structure of the analyzed documents. The advantage of local discourse analyses is their easy applicability in annotation and shallow parsing (looking for surface cues, connectives and other markers), given the usability across different languages and relatively high reliability (e.g. via inter-annotator agreement). As far as language resources aiming at formalized linguistic description of discourse coherence are concerned, internationally, there is a range of corpora annotated for local coherence relations in different languages. They mostly follow the Penn Discourse Treebank (PDTB) annotation style (Prasad et al., 2008, Prasad et al., 2019), compare 2.2. There exists even a multilingual corpus with local discourse annotations in the PDTB style of parallel texts in six languages (Zeyrek et al., 2019).<sup>1</sup> For Czech, there is a publicly available large corpus of local discourse annotation, the Prague Discourse Treebank 2.0 (PDiT 2.0; Rysová et al., 2016), see Section 3. This resource has been developed by the Prague discourse group since 2009 and thus naturally represents the point of departure for this study.

Global coherence models, on the contrary, proceed top-down, from the text as a communicative whole, and postulate a hierarchically interconnected structure of smaller and larger units. These models usually capture each document as a single continuous graph representation with specific properties and constraints on the relations (e.g. tree-like graphs). This concept has a strong potential in the possibility to demonstrate the composition of smaller blocks of the text, as well as to identify more general and more important text contents and relations between them. For global annotations, there are far less projects, compare Section 2, and so far, there is no such project for Czech data. However, global coherence modeling (often in addition to local coherence models), apart from its straightforward application in automatic coherence evaluation (Feng et al., 2014, Lin et al., 2011) can significantly contribute to other NLP tasks, such as summarization (Zhang, 2011), topic identification (Pons-Porrata et al., 2007), text generation (Kiddon et al., 2016), textual entailment (Hagege and Jacquet, 2014) and others.

In this study, we examine the distributions, properties and mutual settings of local discourse relations in order to reveal possible patterns of higher discourse structuring and signs of global discourse coherence. Our research question is: Is global coherence signaled by the same types of language devices as the local one or can we reveal some differences by studying various phenomena in locally annotated data?

The different types of local discourse relations and their mutual settings that we analyze in this study are represented in Example 1 and Figure 1 which outlines discourse annotation in the Czech

---

1. The most exhaustive list of discourse-annotated corpora of any type was collected within the TextLink initiative and can be accessed from this url: <http://www.textlink.iu.metu.edu.tr/corpus-view>

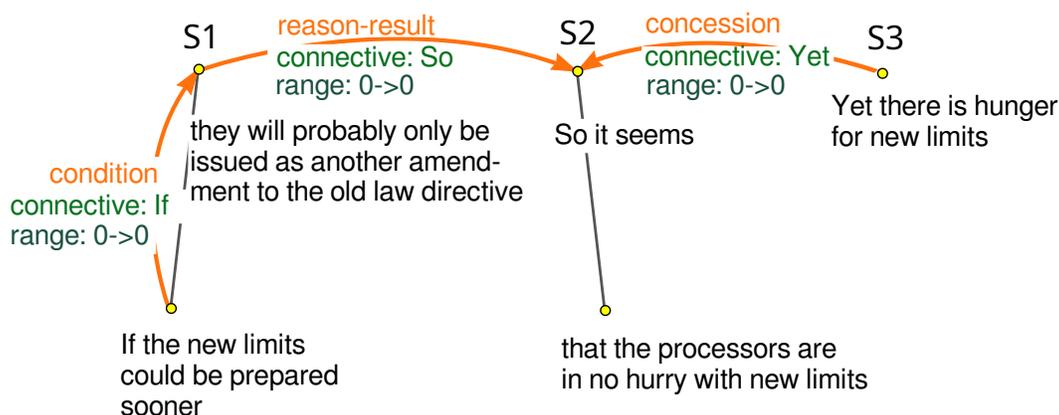


Figure 1: An annotated example of an intra-sentential relation (*condition*), inter-sentential relation within a paragraph (*reason-result*) and a cross-paragraph relation (*concession*) from the Prague Discourse Treebank 2.0. Nodes represent individual clauses, root nodes refer to individual sentences, semantic types of discourse relations are highlighted in orange, connectives in green.

PDiT 2.0. The example demonstrates an *intra-sentential relation*, an *inter-sentential relation within a paragraph* and a *cross-paragraph relation*. Also, a *hierarchical structure* (a nested relation, i.e. a relation fully included within an argument of another relation) is shown by this example text.

- (1) (S1) If the new limits could be prepared sooner, they will probably only be issued as another amendment to the old law directive. (S2) So it seems that the processors are in no hurry with new limits.

[Paragraph boundary]

- (S3) Yet there is hunger for new limits.

Czech original: (S1) Pokud by se podařilo nové limity připravit dříve, budou vydány zřejmě jen jako další novela směrnice starého zákona. (S2) A tak se zdá, že zpracovatelé s novými limity nespěchají.

[Paragraph boundary]

- (S3) Přitom po nových limitech je hlad.

Within the sentence (S1), there is an intra-sentential relation of *condition* signaled by the connective *pokud* [*if*]. Its two arguments are the two clauses in this sentence. The whole sentence (S1) is a left-sided argument of an inter-sentential *reason-result* relation anchored by *a tak* [*so*]. The *if*-relation is a subset of *so*-relation, they are forming a hierarchical structure. These two sentences belong to a single paragraph and the two mentioned discourse relations are local, intra-paragraph relations. Sentence (S3) starts a new paragraph and it is attached to the preceding sentence (S2) by a cross-paragraph inter-sentential *concession* relation with the connective *přitom* [*yet*]. In this case, the cross-paragraph link is local, between adjacent sentences, but in other possible contexts, the left-sided argument of the relation may span across several sentences or/and it can be non-adjacent.

## 1.1 Goals

The main goal of the present study is to systematically analyze existing local discourse annotations of Czech, and partly English, for possible signs of higher/global discourse structure. The analysis shall serve as springboard for a planned global coherence annotation of Czech, similarly as the deep-syntactic layer was informative for the analysis of local coherence relations (Mírovský et al., 2012). An underlying assumption is that even the local annotation already displays such features of global text structure. This assumption is based on our observations from locally annotated data so far: we detected patterns like hierarchical organization of smaller and larger discourse relations, connectives and other discourse cues operating between larger blocks of texts, long-distance relations, genre-related patterns and so on. Also, we hypothesize that in a well-formed, coherent text, the paragraph structure, as a main formal structuring device, must be mirrored in discourse relations and their semantics in some way.

Corpus analyses of possible features of higher/global text structure in this study concern three separate topics:

- The issues of structure, or “the shape” of a text: we investigate mutual configurations of discourse relations (pairwise) and their complexity within locally annotated texts of Prague Discourse Treebank 2.0. We identify and quantify the embeddings, overlappings and crossings of relations, according to a similar study conducted by Lee et al. in 2006 on complexity of discourse structure in the Penn Discourse TreeBank 2.0. Moreover, we look into hierarchies local discourse relations can form. This research was previously published in Poláková and Mírovský (2020). In this new version, we reflect on some feedback by our readers, explain our previous findings in more detail and by more examples, give original Czech annotated examples where they were originally left out due to space limitations, and, most importantly, extend our analysis of hierarchies of nested relations by an analogous one for the English annotations in the PDTB 3.
- The analysis of paragraph-initial connectives, relations they anchor and their respective discourse types (senses). This analysis is intended to reveal any possible differences between (the majority of) relations within individual paragraphs and those cross-paragraph relations in our data that have an explicit marking by a connective.
- The analysis of large arguments, more precisely, of relations with one or both arguments larger than a single sentence, and their properties.

The article is structured as follows. In the second part of the Introduction, we specify the extent of our study and define types of discourse relations with respect to their scope in the text structure. Related research is mentioned in Section 2. Section 3 describes the data format and application framework used in our study and presents the main data resources. Section 4 with four subsections represents the core of the research and offers results of our study in four directions – (i) configurations of pairs of discourse relations, (ii) hierarchies of nested discourse relations, (iii) paragraph-initial senses and connectives, and (iv) relations with large arguments. Section 5 summarizes our findings and offers some perspectives. Examples of texts with deep hierarchies of discourse relations can be found in two Appendices.

## 1.2 Theoretical Aspects and Definitions

For the descriptive purposes of this study, we need to terminologically distinguish several types of coherence. *Global coherence* of a text refers to the coherence of a given document as a whole, including all inner structure of its coherence relations of smaller and greater units. This inner structure is assumed to form hierarchies of smaller and larger relations. On the contrary, *local coherence* refers to the “flat”, chain-like coherence of two minimal subsequent text units, as defined most elaborately in the PDTB 2 (Prasad et al., 2008) approach mentioned above. Further, we distinguish coherence on a higher level, or the so-called *higher text structure*, which describes coherence relations between/among larger text blocks, but not necessarily in a whole document. Typically, a distinctive higher structure in a text is formed by (multi-sentential) lists and enumerations and it often relates to genre rules or practises (e.g. in legal texts, instructions, recipes). Independently, we also distinguish *inter- or cross-paragraph coherence*, that means, coherence relations between individual paragraphs as integral units but also between smaller units belonging to two different paragraphs. Coherence relations within individual paragraphs are then again referred to as *local relations*, and they can be either *intra- or inter-sentential*, or belong to a higher structure, if made up of larger units (more sentences).

Obviously, these categories are not rigorous and mutually exclusive, as already the very purpose of this study – looking for global features in a locally annotated data – suggests. This gives rise to a question, e.g., how to approach the relation of the last sentence of one paragraph to the immediately following first sentence of the next paragraph, or, how to treat intra-sentential lists spanning across several paragraphs. We will address these and further related issues further in Section 4.

A note should be made to the use of the term *hierarchy* in this study. Whereas we explore the hierarchical organization of discourse *relations* as wholes, there also is another understanding of hierarchy in discourse, in terms of the ways discourse *units* organize themselves based on criteria such as their relative importance (nuclearity in Rhetorical Structure Theory). We do not address the latter here.

## 2. Related Research

From the wide range of coherence theories which address a text/document as a structured whole and offer some type of a formalized representation of this whole (e.g. Grosz and Sidner, 1986, Hobbs, 1979, Asher and Lascarides, 2003, Wolf and Gibson, 2005),<sup>2</sup> our approach is methodologically closest to the Penn Discourse Treebank and Rhetorical Structure Theory (RST), which are introduced in more detail in the following sections.

### 2.1 Rhetorical Structure Theory

The Rhetorical Structure Theory (RST, Mann and Thompson, 1988) is one of the most influential frameworks among the global coherence models. It was originally developed with the intention to model text coherence in order to study computer-based text generation. The main principle of RST is the assumption that coherent texts consist of minimal units, which are linked to each other, recursively, through rhetorical relations and that coherent texts do not exhibit gaps or non-sequiturs (Taboada and Mann, 2006). The RST represents the whole text document as a single (projective)

---

2. For a detailed classification of discourse approaches within the local/global space of coherence relations see Bateman and Rondhuis (1997).

tree structure. Basic features of these structures are the *rhetorical relations* between two textual units (smaller or larger blocks that are in the vast majority of cases adjacent) and the notion of *nuclearity*. For the classification of RST rhetorical relations, a set of labels was developed, which originally contained 24 relations, but the authors themselves add that it is an open set “susceptible to extension and modification for the purposes of particular genres and cultural styles” (Mann and Thompson 1988, p. 250). The type of a rhetorical relation is defined with respect to the author’s intended effect on the reader together with the application of the principles of nuclearity.

The RST has gained great attention, it was further developed and tested, language corpora were built with RST-like discourse annotation, such as for English the RST Discourse Treebank (RST-DT, Carlson et al., 2003) and its extension RST-Signalling Corpus (Das et al., 2015), for German (Stede and Neumann, 2014), Spanish (Da Cunha et al., 2011), Basque (Iruskieta et al., 2013) and also for some other languages.

On the other hand, the framework was repeatedly criticized with regard to some of its theoretical claims, above all, concerning the question of adequacy/sufficiency of representation of a discourse structure as a tree graph.<sup>3</sup> Linguistically, the strong constraints of a tree (no crossing edges, one root, all the units interconnected etc.) gave rise to a search for counter-examples in real-world texts. It was shown that not only adjacent text units exhibit coherence links and that there are even cue phrases, which connect non-adjacent units and thus support the claim that a tree graph is too restricted a structure for an adequate discourse representation (Wolf and Gibson, 2005). Therefore, more complex graphs with crossings and overlaps should be adequate, which resulted in the creation of the Discourse GraphBank (Wolf et al., 2005), a resource of the same texts annotated with the main diverging principle of relaxing the tree-ness constraint.

On these grounds and in the same way like Lee et al. (2006), we try to demonstrate where the local and global analytic perspectives meet and interact. Our analysis of Czech data thus contributes new empirical material to the scientific debate on whether projective trees are descriptively adequate to represent structure of texts (e.g. Marcu, 2000, Wolf and Gibson, 2005, Lee et al., 2008, Egg and Redeker, 2010).

Additionally, there have been some RST-based studies investigating the distributions of rhetorical relations on different levels of text, e.g. Williams and Reiter (2003), Liu and Zhang (2016). Their research questions are similar to ours in the section on paragraph-initial relations and cues 4.3. The latter study introduces a RST-tree conversion to dependency trees with different levels of granularity, with a node representing a clause, a sentence or a paragraph. They point out that most rhetorical relations in the RST-DT treebank occur on all levels but with different percentages and the two upper levels of discourse processes are more alike in many aspects, which are observations that also follow from our local data analyses (4.3).

## 2.2 Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) represents a lexically based, local model of discourse. Its analysis of discourse relations consists primarily in finding and analyzing lexical cues of discourse coherence as “anchors” of discourse relations. Such a cue, a discourse connective, is defined as a discourse-level predicate opening positions for two discourse arguments. Discourse connectives include coordinating conjunctions (e.g. *and*, *but*), subordinating conjunctions (*because*, *if*) and

---

3. Basically a constituency tree, which is in its nature projective and does not allow crossing edges, in comparison to the basic mathematical definition of a tree graph.

discourse adverbials (*nevertheless*). Apart from connectives, the two discourse arguments of a discourse relation (and their extent) and the semantic type (sense) of a discourse annotation were annotated. In 2004, the first version of Penn Discourse Treebank was released (Miltsakaki et al., 2004). The second release of the PDTB four years later includes annotation of the ca. 49,000 sentences of the Wall Street Journal part of the Penn Treebank (Prasad et al., 2008). Apart from explicit connectives, other phenomena have been annotated in this version, mainly implicit relations and attribution (attributing beliefs and assertions to agents making them) alternative lexicalization of a connective (AltLex, e. g. *that is why*), entity based relations (EntRel) and places with weak coherence (NoRel). In its latest version PDTB 3.0 (Prasad et al., 2018, Prasad et al., 2019), the annotations were enriched by many relations mostly in intra-sentential contexts, the sense taxonomy was revised and the existing annotations enhanced in many aspects.

The PDTB-style connective/argument analysis has become very popular, also because such an analysis requires less interpretation and pragmatic inference than the RST analysis. The PDTB authors also claim that their approach is theory-neutral, independent from any syntactic theory, and as such can be transferred to other languages.

In the first part of our analysis on configurations of positionally close relation pairs (4.1), we relate our findings to a similar research conducted on the second version of the English Penn Discourse Treebank in Lee et al. (2006). The aim of their study was to describe and quantify the configurations of discourse relations as typical or less typical in terms of discourse complexity. The complexity of discourse configurations there is being compared to the complexity of relations in syntax, but also refers to the principles of the global Rhetorical Structure Theory, in particular the representation of any text document as a single tree-like structure with strong constraints. The study actually tries to answer the question whether their empirical locally annotated data would fulfill or violate these strong constraints. Therefore Lee et al. (2006) studied various types of overlaps of discourse relations. They encountered a variety of patterns between pairs of discourse relations, including nested (hierarchical), crossed and other non-tree-like configurations. Nevertheless, they conclude that the types of discourse dependencies are highly restricted since the more complex cases like crossings and partial overlaps can be factored out by appealing to discourse notions like anaphora (non-structural, possibly long-distance tie) and attribution (attribution span as a known mismatch between the syntactic and discourse structure), and argue that pure crossing dependencies, partially overlapping arguments and a subset of structures containing a properly contained argument should not be considered part of the discourse structure. The authors challenge Czech discourse researchers to introduce a similar study (footnote 1 in their paper, *ibid.*) in order to observe and compare the complexity of discourse and syntax dependencies in two typologically different languages.<sup>4</sup>

### 3. Data and Tools

Our study is based primarily on two data resources: for Czech, the Prague Discourse Treebank 2.0 (PDiT; Rysová et al., 2016, Zikánová et al., 2015),<sup>5</sup> and, for English, the Penn Discourse Tree-

4. From this viewpoint, we do not assume substantial differences between the two languages in discourse structure itself. However, there are surely larger differences on “lower” levels of linguistic description, in our case most visibly connective repertoires and syntactic properties of the two languages, which can in the outcome influence some motivations in the annotation design and thus effect the resulting discourse structure, compare Mírovský et al. (in prep.).

5. Please note that the discourse annotation in PDiT 2.0 is the same as in PDT 3.5 (Hajič et al., 2018) and the upcoming PDT-C 1.0, which are newer versions of the PDiT 2.0 data.

CONTRAST	EXPANSION
<i>confrontation</i>	<i>conjunction</i>
<i>opposition</i>	<i>conjunctive alternative</i>
<i>restrictive opposition</i>	<i>disjunctive alternative</i>
<i>pragmatic contrast</i>	<i>instantiation</i>
<i>concession</i>	<i>specification</i>
<i>correction</i>	<i>equivalence</i>
<i>gradation</i>	<i>generalization</i>
CONTINGENCY	TEMPORAL
<i>reason–result</i>	<i>synchrony</i>
<i>pragmatic reason–result</i>	<i>precedence–succession</i>
<i>explication</i>	
<i>condition</i>	
<i>pragmatic condition</i>	
<i>purpose</i>	

Table 1: Semantic types of discourse relations in the PDiT 2.0

bank ver. 3.0 (PDTB; Prasad et al., 2019). These two corpora are comparable in size (approx. 50 thousand sentences each), genre (journalistic texts) and they have both been manually annotated for discourse relations in comparable approaches (PDTB and the PDTB-like modification in the PDiT). Namely, for explicit relations, all connectives in a given text were identified, their two arguments were detected and a semantic/pragmatic relation<sup>6</sup> between them was assigned to the relation. Discourse connectives in PDiT 2.0 are of two types: primary and secondary, according to Rysová and Rysová (2014). Primary connectives are defined as grammaticalized expressions such as *because* or *therefore* whereas secondary connectives are not (yet) fully grammaticalized structures such as *except for this, the reason was* or *for this reason*. The notion of secondary connectives in Prague Treebank roughly corresponds to the categories of AltLex and AltlexC (alternative lexicalization tokens and alternative lexicalization constructions) in PDTB 3.

The PDTB and PDiT use similar sets of discourse relation types – the set of discourse types in PDiT is inspired by the Penn Discourse Treebank 2.0 sense hierarchy (Prasad et al., 2008), see the complete list in Table 1. The taxonomy is two-level, but in the annotation only the second-level relations can be assigned. Confrontation in the Czech taxonomy equals to the original juxtaposition in the PDTB 2 (mostly represented by connectives like *while, whereas*), restrictive opposition also includes exception (*only, except that...*), correction equals to substitution (*not A, but B, instead*), explication is giving evidence by other means than reason-result.<sup>7</sup> Pragmatic labels are used in accordance with the PDTB 2 pragmatic labels.

6. "sense" in the PDTB terminology, "discourse type" in PDiT 2.0

7. strongly represented by a typical Czech connective *totiž* (*actually, since, I mean, as a matter of fact*)

PDiT 2.0 contains 21 223 explicit discourse relations (i.e. relations signalled by explicit discourse connectives, both primary and secondary),<sup>8</sup> while the PDTB 3 contains 25 696 such relations (we take into account relations of type Explicit, (AltLex) and AltLexC).<sup>9</sup>

Both PDiT and the PDTB are available in the Prague Markup Language format (PML; Pajas and Štěpánek, 2008),<sup>10</sup> an XML-based format and application framework designed for multi-layer linguistic annotations with available tools allowing for complex linguistic studies over PML data: btred for scripting in Perl and Prague Markup Language - Tree Query (PML-TQ; Pajas and Štěpánek, 2009) as a graphically oriented querying system.

## 4. Analysis

This section is divided into four subsections according to the research directions in which we address different aspects of a higher/global text structure traceable in locally annotated corpora. 4.1 – configurations of pairs of discourse relations, a study carried out on the basis of a similar research by the PDTB group and, like this study, contributing to the debate on acceptable “shapes” of an interconnected coherence representation of a text; 4.2 – hierarchies of nested discourse relations – a study that reveals if local discourse relations, without having the underlying assumption of hierarchy, also exhibit some type of hierarchical structure; 4.3 – paragraph-initial senses and connectives, a survey on cross-paragraph coherence in comparison to intra-paragraph settings; and 4.4 – relations with large arguments, as these are rather unexpected/unusual in local settings and can denote possible non-elementary units of higher text structuring.<sup>11</sup>

### 4.1 Configurations of Close Relation Pairs

For this part of the analysis, and also for the subsequent one described in 4.2, we take advantage of the fact that the corpora used in the PDTB and our studies are comparable in many aspects (as mentioned earlier) and also contain a similar number of annotated discourse relations (approx. 20 thousand). However, it is important to notice that Lee et al. (2006) included also implicit relations in their study and conducted the research on the older version of the PDTB (2.0). The Lee et al. (2006) study defines six basic types of patterns of relation pairs: independent relations, full embeddings, shared arguments, properly contained arguments, pure crossings and partially overlapping arguments. In our analysis, we decided to make a more fine-grained classification of the patterns to cover all possible settings and to get more detailed insight into the configurations, and ended up with 17 categories, compare Table 2. Sometimes we also use different pattern naming (e.g. *independent relations* vs. *adjacency*), but for those patterns accounted for in both studies, we state how the names map.

In our data, we have collected and classified types of relative positions of neighbouring discourse relations in the corpus, using a scripting tool btred as a part of an application framework for PDiT-

8. We exclude the 361 list relations from this study, i.e. relations between subsequent members of enumerative structures. They represent a special text structure on their own, and as such, in the Prague approach they stand aside the binary discourse relations and discourse type taxonomy.

9. Implicit discourse relations have been annotated on a part of the PDiT 2.0, see Zikánová et al. (2019); this annotation was not taken into account in the present analysis, as the size of the implicit datasets is different at present.

10. The PML is the primary publication format for PDiT. The English PDTB was first transformed to the PML format, for the purposes of studying discourse annotations in a unified format.

11. The first and the second subsections/topics are closely related, as hierarchies are a subset of possible relation configurations (nested relations pattern), but we keep them separate in the study.

Line	Frequency in PDiT	Pattern	Visualization
1	6 572 (983) 37.3% (32.9%)	adjacency	<-----> <-----> <-----> <----->
2	1 923 (865) 10.9% (29.0%)	progress (shared argument)	<-----> <-----> <-----> <----->
3	51 (21) 0.3% (0.7%)	total overlap	<-----> <-----> <-----> <----->
4	266 (116) 1.5% (3.9%)	left overlap right adjacency	<-----> <-----> <-----> <----->
5	25 (10) 0.1% (0.3%)	left overlap right contained	<-----> <-----> <-----> <====>
6	31 (10) 0.2% (0.3%)	right overlap left adjacency	<-----> <-----> <-----> <----->
7	17 (8) 0.1% (0.3%)	right overlap left contained	<-----> <-----> <====> <----->
8	250 (146) 1.4% (4.9%)	containment I	<-----> <-----> <====> <----->
9	190 (150) 1.1% (5.0%)	containment II (opposite)	<-----> <----> <-----> <----->
10	83 (14) 0.5% (0.5%)	both args contained I	<-----> <-----> <====> <====>
11	73 (8) 0.4% (0.3%)	both args contained II	<-----> <----> <====> <----->
12	3 591 (299) 20.4% (10.0%)	left hierarchy	<-----> <-----> <====> <====>
13	3 543 (56) 20.1% (1.9%)	right hierarchy	<-----> <-----> <====> <====>
14	10 (10) 0.1% (0.3%)	crossing	<-----> <-----> <-----> <----->
15	883 (196) 5.0% (6.6%)	envelopment	<-----> <-----> <-----> <----->
16	3 (2) 0.0% (0.1%)	partial overlap	<-----> <-----> <-----> <----->
17	8 (5) 0.0% (0.2%)	partial overlap II	<-----> <-----> <-----> <----->

Table 2: Patterns of adjacent, overlapping or embedded pairs of explicit discourse relations in PDiT 2.0. The column visualization shows relative position of the arguments read from left to right; arguments of one relation are marked with '- ', the other one with '= '. The total number of such close relations in PDiT was 17 628; 109 of them (0.6%) did not fit the listed patterns. Numbers in brackets mean frequencies if only inter-sentential discourse relations (arguments in different sentences) were taken into account; in total, there were 2 984 such close relations in PDiT, 85 of them (2.8%) did not fit the listed patterns.

types of corpora (Pajas and Štěpánek, 2008). We define a *close relation pair* as a pair of discourse relations that are either adjacent (the left argument of one relation immediately follows the right

argument of the other relation), or they overlap in one of many possible patterns. We first analyze the detected patterns in PDiT and, next, a specific subsection is devoted to the description of the detected hierarchical structures (4.2).

We were able to detect 17 628 close relation pairs in PDiT, and for each such pair, we investigated its pattern, the mutual arrangement of the two relations. In Table 2, these patterns are also graphically illustrated. The table shows figures for all explicit relation pairs and, in brackets, for inter-sentential relation pairs only.

The most common settings for close relations in PDiT are **pure adjacency** (succession) of two relations (line 1: 6 572 cases in PDiT), and “full embeddings”, in other words **two-level hierarchies**, in total 7 134 in PDiT (lines 12 and 13). These configurations (lines 1, 12 and 13) represent together slightly more than 3/4 of all detected patterns in PDiT. They are also referred to as *very “normal” structural relationships* in Lee et al. (2006, p. 82).

The next-largest group is **progress** (line 2), a shared argument in the PDTB terminology, with 1 923 instances or 10.9% of all patterns. Lee et. al. report 7.5% of this type, which is fairly comparable.

**Total overlap** (line 3) is caused by the possibility to annotate two different relations between the same segments for co-occurring connectives, as in *because for example* or *but later*. It occurs 51 times (0.3%) in PDiT. This pattern may vary a lot in different annotation schemes, as there may be different approaches to handling co-occurring connectives, or even the possibility of a co-occurrence of an explicit and an implicit relations between the same segments, which is the approach taken by some newer corpora, e.g. by the PDTB in its version 3.

The **envelopment** (line 15) concerns in the vast majority of cases a non-adjacent (long-distance) relation and another relation placed in the text between the arguments of the non-adjacent relation. The enveloped relation is often a sentence with an inner syntactic structure annotated. The same is often true for hierarchical patterns (lines 12 and 13) and it explains the big difference observed in both corpora between all detected envelopment and hierarchical patterns and only the inter-sentential ones. Linguistically, some of the envelopment cases are sentences headed by two attribution spans (verbs of saying) and some structure in the reported content in between, also cases of two linked reporter’s questions in an interview and the inner structuring of the interviewee’s answer, but also texts with no striking structural reasons for such an arrangement. In PDiT, **envelopment** represents ca. 5% of all settings and is certainly worth of further investigation.

Patterns with **properly contained arguments**, either one of the arguments (lines 5, 7, 8 and 9) or both (10 and 11), very often involve “skipping one level” in the syntactic tree of a sentence, see Example 2<sup>12</sup> of the type 8 (containment I), i.e. exclusion of a governing clause from the argument (mostly an attribution span), that makes its syntactically dependent clause (mostly a “reported-content-argument”) to a subset of an other argument of the other relation represented (mostly) by a whole sentence. In Example 2, the text span *a young researcher works with enthusiasm for science, regardless of salary* is a left-sided argument of the *but*-relation (in bold) and, at the same time a subset, properly contained, in a larger right-sided argument of the *therefore*-relation (in italics). The governing clause *It is therefore appropriate... the fact that* is thus the mismatch, it represents the difference between proper containment and the much more frequent shared argument pattern (line 2).

---

12. From now on in the annotated examples, relation 1 is highlighted in italics, relation 2 in bold. The connectives are underlined.

- (2) *The gap in the standard of living that appears between the qualified scientific elite and the business sphere, right now, at the beginning of the transformation of the society, will leave traces. It is therefore appropriate to pamper young researchers and not misuse the fact that a young researcher works with enthusiasm for science, regardless of salary. **But a person who begins to find his mission in research also starts a family, wants to live at a good place and live with dignity.***

Czech original: *Mezera na úrovni životního standardu, která se objevuje mezi kvalifikovanou vědeckou špičkou a sférou podnikatelskou, právě teď, v počátcích transformace společnosti, přece jen zanechá stopy. Je tedy namístě hýčkat dorost a nehřešit na to, že mladý badatel pracuje s nadšením pro vědu, bez ohledu na plat. **Jenže člověk, který začíná nalézat své poslání ve výzkumné práci, také zakládá rodinu, chce bydlet a důstojně žít.***

Besides the discussed attribution (introductory statements) that has been excluded from the argument, this mismatch in the shared argument extent is also the case of the annotation of some secondary connectives realized by whole clauses (like *This means that...*, Example 3, type 9, containment II). These verb phrases are not treated as parts of any of the arguments they relate to and pose a methodological issue. In the representation in Example 3, the clause *This means that* is not in italics, it is not considered to be a part of any argument of the left relation, i. e. the relation it anchors.

- (3) *This brief overview essentially exhausts the areas of notarial activities within the framework of free competition between notaries. **This means that in these notarial agendas, the client has the option of unlimited choice of notary at his own discretion, as the notary is not bound to the place of his work when providing these services.***

CZ: *Tímto stručným přehledem jsou v podstatě vyčerpány oblasti notářských činností v rámci volné konkurence mezi notáři. **To znamená, že v těchto notářských agendách má klient možnost neomezeného výběru notáře podle vlastní úvahy, neboť notář při poskytování těchto služeb není vázán na místo svého působení.***

A third setting concerns multi-sentence arguments, where the contained argument is typically a single sentence. Patterns with properly contained arguments (lines 5, 7 – 11) represent in total 3.6% (638) of all patterns in PDiT.

A (pure) **crossing** is a setting where the left-sided argument of the right relation comes in between the two arguments of the left relation, compare line 14 in Table 2. Pure crossings violate the RST constraints most visibly, with crossing edges, so the debate on tree adequacy often circles around the acceptability of crossings in discourse analysis. Lee et al. (2006) identify 24 cases (0.12%) of crossings in the PDTB2. We detected only 10 such cases in PDiT, which is a very small proportion. Manual inspection of the cases of crossing revealed several different scenarios, from clearly incorrect annotation, more interpretations possible, across cases with attribution spans in between, to a few, in our opinion, perfectly sound analyses, as exemplified by Example 4 from PDiT.

- (4) (a) What can owners and tenants expect, what should they prepare for?  
 (b) *The new legislation should allow all owners to sell apartments.* (c) **It is most urgent for flats owned by municipalities, as they manage about a quarter of the housing stock of the Czech Republic and some are - mainly for financial reasons - interested in mon-**

**etizing a part of their apartments.** (d) *The law should also allow to complete transfers of housing association apartments and their sale to its members.* (e) **This is also not a “small portion”, but a fifth of the total number of dwellings.**<sup>13</sup>

CZ: (a) S čím mohou vlastníci i nájemci počítat, na co by se měli připravit?

(b) *Nová právní úprava by měla umožnit všem vlastníkům prodávat byty.* (c) **Nejnáléhavěji to vystupuje do popředí u bytů ve vlastnictví obcí, protože ty spravují asi čtvrtinu bytového fondu České republiky a některé mají - především z finančních důvodů - zájem část svých bytů zpeněžit.** (d) *Zákon by měl také umožnit dokončení převodu i možnost prodeje družstevních bytů členům.* (e) **Ani zde nejde o “malou porci”, ale o pětinu celkového počtu bytů.**

If we accepted the possibility that not only (b), but a larger (b+c) unit relates to (d) in the *also*-relation, which would be a completely fine interpretation in the Prague annotation, the relation of (e) – the *neither*-relation – in our view still cannot accept just (d) as its left-sided argument. We also think this case cannot be factored out due to anaphora. There is, for sure, room for different interpretations within different theories, we just offer our data, state our view and admit that crossing structures are extremely rare even in our empirical data.

**Partial overlap** is a type of structure that violates the RST tree constraint, too. Lee et al. (2006) could only find 4 such cases. We detected 11 cases in PDiT (lines 16 and 17 of Table 2). They often include large arguments of untypical range (2.5 sentences etc.) which can be questioned. Some of the relations also include secondary connectives with strong anaphoric links (*in this respect, given the fact that etc.*). These relations can be factored out, yet, again, even among the small number of cases in PDiT there were linguistically acceptable ones, compare Example 5 (partial overlap II).

- (5) The responsibility of the future tenant of this 103,000-square-metre area will be to care for all properties, including their maintenance and repairs. *The tenant will also have to resolve the parking conditions for market visitors and to meet the conditions of the Prague Heritage Institute during construction changes due to the fact that the complex is a cultural monument. The capital city at the same time envisages preserving the character of the Holešovice market.*

CZ: Povinností budoucího nájemce tohoto areálu o rozloze 103 tisíc metrů čtverečních bude mj. péče o všechny nemovitosti včetně jejich údržby a oprav. *Nájemce bude také muset vyřešit podmínky parkování pro návštěvníky tržnice a splnit podmínky Pražského ústavu památkové péče při úpravách objektů vzhledem k tomu, že jde o kulturní památku. Hlavní město přítom počítá se zachováním charakteru holešovické tržnice.*

## 4.2 Hierarchies

The purpose of looking for hierarchical structures in the locally annotated data is to discover to what extent such an annotation shows signs of some higher (global) structure, too. We do not claim that the trees detected by us are the trees a global analysis like RST would discover, but we demonstrate the existence of some hierarchical text structure in local annotation. Some of it could perhaps partially match to RST-formed subtrees (and definitely there would be an intersection of separate relations, compare e.g. intersections in Wall Street Journal local and global annotations in

13. The “also-not” connective is originally in Czech *ani*, in the meaning of *neither*. Lit. translation: “Neither here is concerned a small portion...”

Line	Frequency	Depth	Hierarchy
1	381 (10)	3	A ( B ( C ))
2	64 (1)	3	A ( B ( C ) D )
3	50	3	A ( B C ( D ))
4	23 (1)	3	A ( B ( C ) D ( E ))
5	20	3	A ( B ( C D ))
6	0 (1)	3	A ( B ( C ) D E )
7	0 (1)	3	A ( B C ( D ) E )
8	19	4	A ( B ( C ( D )))
9	7	4	A ( B ( C ( D )) E )
10	5	4	A ( B C ( D ( E )))
11	5	4	A ( B ( C ( D )) E ( F ( G )))
		...	
12	2	5	A ( B ( C ( D ( E ))) F )
13	1	5	A ( B C ( D ( E ( F ))) G )
14	1	5	A ( B ( C ( D ) E ( F ( G ))) H I J ( K ))
15	1	5	A ( B ( C ( D ( E )))
16	1	5	A ( B ( C ( D E ( F )))

Table 3: Selected schemes of hierarchies of explicit discourse relations in PDiT. Numbers in brackets mean frequencies of hierarchy schemes if only inter-sentential discourse relations were taken into account (no other inter-sentential hierarchies were encountered in the data). The three dots in the middle of the table indicate that there can be more cases of different 4-level hierarchy patterns.

Poláková et al., 2017), but this is yet to be investigated. We are also aware, as pointed out in Egg and Redeker (2010), that minimal, local annotations do not normally form a connected graph.<sup>14</sup>

#### 4.2.1 HIERARCHIES IN THE PDiT 2.0

For the study of hierarchical structures, we used pairs of nested relations where one of the relations is as a whole included in one argument of the other relation (patterns corresponding to lines 12 and 13 from Table 2) to recursively construct tree structures out of pairs of the nested relations. Based on the quantitative results, we inspected selected samples of the detected patterns manually, in order to check the script outcome and to provide a linguistic description and comparison. The results on the whole PDiT data are displayed in Table 3, arranged according to the scheme of such hierarchy trees (identical structures are summed and represented by the hierarchy scheme). We only mention cases where there are at least three levels in the tree, as two-level hierarchies are part of Table 2.

For explanation: The scheme “A ( B )” means that the whole relation B is included in one of the arguments of the relation A (this is, of course, only a two-level tree). The scheme “A ( B C ( D ))”

14. And the more so, as we do not include implicit and entity-based relations into our study.

means that relations B and C are all included in the individual arguments of relation A (without specifying in which argument they are, so they can be both in one argument or each in a different argument) and the relation D is completely included in one of the arguments of relation C. It is a three-level hierarchy. Generally, we count the depth (number of levels) of a hierarchy tree as a number of nodes in the longest path from the root to a leaf.

There are many sub-hierarchies in a large/deep hierarchy, for example “B ( C ( D ))” is a sub-hierarchy of “A ( B ( C ( D )) E )”; such sub-hierarchies, however, are not counted in Table 3, i.e., each hierarchy is only counted in the table in its largest and deepest form as it appeared in the PDiT 2.0 data.<sup>15</sup>

In the PDiT data, local discourse relations form hierarchies up to five levels. We have identified 5 patterns of 5-level hierarchies (5-LH), with the total of 6 instances, see Table 3. There is also a number of 3- and 4-level hierarchies. An analysis of random samples (and of all the deepest ones) revealed, surprisingly, that there may be a 4-level hierarchy spanning 11 sentences, but also a 5-LH spanning only two sentences, from which one is typically a more complex compound sentence. The “longest” of the 5-LHs includes also 11 sentences (line 14) and it also exhibits branching (D, G and K as leaves, where the G-path is the deepest). One of the 5-LHs should be in fact one level flatter (line 12), as the lowest two relations are three coordinated clauses with two *and*-connectives: “*the troops protected them and fed them and gave them the impression that they were invulnerable...*”. Such structures are notoriously hard to interpret for any framework, yet in Prague annotation, the annotation is incorrectly hierarchical where it should have been flat.

For a better illustration of the hierarchical configuration of the relations, a text sample with a 4-LH is analyzed in Appendix 1 to this study. It covers one whole paragraph and a part of a preceding one (11 sentences). Its pattern is A ( B ( C ( D ) E ) F G ).<sup>16</sup>

To find out how much structure is involved only within individual sentences, i.e. how much of sentential syntax forms the hierarchies, in a second phase we filtered out all intra-sentential relations. The numbers in brackets give counts for patterns of hierarchies, if only inter-sentential relations (i.e. arguments in different sentences) are accounted for.<sup>17</sup> The hierarchies of this type are much less frequent and their maximum depth is just 3, which implies that beyond the sentence boundary, local annotation of explicit connectives does not represent hierarchical text structuring very often in the PDiT 2.0.

#### 4.2.2 HIERARCHIES IN THE PENN DISCOURSE TREEBANK 3.0

Having at our disposal the Penn Discourse Treebank 3.0 annotations converted into the Prague PML format (Poláková et al., 2017), we can use the same procedure to search for hierarchical structures also in these English locally annotated data. We take into account relations of the type Explicit, AltLex and AltLexC. The results are summed up in Table 4. Similarly as in the experiment with the Czech data, we have detected hierarchies of relations up to 5 depth levels. Overall, their counts are

15. This also explains the zeros in Table 3. The empty line in the table suggests that there are more different patterns of hierarchies of the given depth, the same holds also for Table 4.

16. We do not present a 5-LH here, as the largest one is too complex and the smaller ones include two sentences only, so the main structure is syntactic in nature.

17. Please note that hierarchies counted in brackets (without intra-sentential relations, column 3 in Tables 3 and 4) are not a subset of the respective hierarchies from the same table row that also include intra-sentential relations – adding intra-sentential relations into the hierarchies means that a particular hierarchy pattern may change (be enlarged). This happens most clearly in lines 6 and 7 of Table 3.

Line	Frequency	Depth	Hierarchy
1	232 (2)	3	A ( B ( C ))
2	63 (2)	3	A ( B ( C ) D )
3	46 (2)	3	A ( B C ( D ))
4	20	3	A ( B ( C D ))
5	10	3	A ( B ( C ) D ( E ))
6	8	3	A ( B ( C ) D E )
	...		
7	2 (1)	3	A ( B ( C ) D E F )
	...		
8	5	4	A ( B ( C ( D )) E )
9	3	4	A ( B ( C ( D )))
10	2	4	A ( B ( C ( D ) E ) F )
11	1	4	A ( B C D E F ( G ( H )))
12	1	4	A ( B ( C D ( E ) F ) G ( H ) I ( J ) K ( L ))
	...		
13	1	5	A ( B C D ( E ( F ( G ))))

Table 4: Selected schemes of hierarchies of explicit discourse relations in the PDTB 3.0. Numbers in brackets mean frequencies of hierarchy schemes if only inter-sentential discourse relations were taken into account (no other inter-sentential at least 3-level hierarchies were encountered in the data).

slightly lower, e.g. the simplest 3-LH pattern (A (B (C))) is 232 in comparison to 381 in the Czech data. The maximum depth is also 5 but there is only one such detected hierarchy, compare the last line of Table 4. The source text with all the relations annotated within this hierarchy is presented in detail in the Appendix 2 to this study. The hierarchy spans across 2 paragraphs (7 sentences), and its pattern is A ( B C D ( E ( F ( G )))), which implies that the relations B and C do not take part in the deepest branch of the hierarchy, they are only included in the highest relation A. Further, it can be observed that the two lowest relations, G (the deepest one) and F, are intra-sentential, while the hierarchy crosses the sentence boundary with its relation E, which is the lowest inter-sentential relation. So, only the relations A and D, the arguments of which span across two or more sentences<sup>18</sup> in our understanding contribute in a way to a higher discourse structure.

The number of hierarchies formed by inter-sentential relations only is even lower in the PDTB 3, only 7 (compared to 14 in the PDiT 2.0), with the same maximum depth of 3 levels. A hypothesis for the relatively small number of hierarchies built by only inter-sentential relations in both corpora is that only some of the connectives operating at higher discourse levels were identified and annotated as such, some of them were assigned local coherence links due to the minimality principle.<sup>19</sup>

18. more precisely, one argument in each of them spans two or more sentences

19. The minimality principle instructs the annotators to mark as an argument as many clauses and/or sentences as are *minimally required* and *sufficient* for the interpretation of the relation. It was applied both in PDTB and in PDiT

Number of documents	3 165
Number of sentences	49 431
Number of sentences without headings, captions and metatext	44 979
Number of paragraphs	14 790
Number of paragraphs without headings etc.	11 643
The average number of sentences per paragraph <sup>21</sup>	3.9

Table 5: Basic properties of the inner structure of texts in the Prague Discourse Treebank

This issue was recently discussed in Poláková and Mírovský (2019) with focus on paragraph-initial connectives and we dedicate the following subsection to this topic.

### 4.3 Paragraph-initial Semantic Types and Connectives

A possible different scope of connectives in paragraph-initial positions was previously indicated in a study analyzing discourse connectives with anaphoric properties and their ability to relate to a distant, non-adjacent left-sided argument (Poláková and Mírovský, 2019). A special set of cases was defined where the connective actually does not relate to a short non-adjacent segment on its left, but to an adjacent, quite larger segment of text and is interpretable as a means of higher discourse structuring. From another perspective on paragraph boundary, cross-paragraph discourse relations were recently investigated in Prasad et al. (2017) for implicit relations, leading to the observation that a first sentence in a given paragraph semantically relates to the immediately preceding last sentence of the previous paragraph in only 52% of their sample, with 48% having links to non-adjacent left contexts.

Our analysis of the features of cross-paragraph coherence only takes explicit relations into account (including secondary connectives), although we acknowledge that implicit connecting or other signalling is common between paragraphs. We focus on the following issues: (i) Is there a difference between the semantics of discourse relations that ensure continuity between paragraphs on one side, and relations within an individual paragraph on the other? Are certain relations more typical for cross-paragraph coherence, while others are typically local? (ii) Are discourse connectives in the cross-paragraph usage different from locally used connectives?

Because of limitations of the actual version of the conversion of the PDTB data to the PML format, in this subsection and also in the study of paragraph-initial connectives in the subsequent subsection we only examined texts of the Prague Discourse Treebank 2.0 (as characterized in Table 5), via PML-TQ queries and using information about paragraph numbers available at roots of the deep-syntactic trees.<sup>20</sup>

In terms of the higher structure and local relations in a text, three sets of relations can be distinguished. First of them are the relations between paragraphs, or cross-paragraph relations, as the top set (1). The rest, i.e. intra-paragraph relations, can be divided into inter-sentential (2) and intra-

annotations. The PDTB moreover annotates supplementary material to an argument, where needed (Prasad et al., 2008).

20. Information about paragraph numbers in a direct form of attribute `para_no` is only available in the PDiT data accessible via the PML-TQ search engine and is taken from identifiers (attribute `id`) of t-roots – roots of the tectogrammatical trees.

21. The average length of a paragraph is counted after the exclusion of headings, captions and metatext, i.e. 44 979 / 11 643.

(1) Cross-par inter-S		(2) Intra-par inter-S		(3) Intra-S	
conjunction	236	opposition	1 584	conjunction	6 154
opposition	207	conjunction	1 322	reason-result	1 663
reason-result	125	reason-result	1 056	condition	1 400
confrontation	44	confrontation	273	opposition	1 388
generalization	43	concession	252	concession	633
specification	42	precedence	240	precedence	592
concession	32	gradation	197	specification	521
instantiation	22	explication	155	purpose	412
precedence	20	restr. opposition	153	confrontation	348
restr. opposition	19	correction	127	correction	323

Table 6: 10 most common semantic types of explicit discourse relations for each dataset (number of occurrences)

sentential ones (3); the differences between these groups are mainly due to the syntactic structure. In our analysis, we want to focus specifically on the relations between paragraphs (1). These relations do not reflect syntactic structure that much, which is why we compare them mainly to the separate group of intra-paragraph inter-sentential relations (2). However, in order to keep the picture of discourse relations complete, we also include a comparison with (intra-paragraph) intra-sentential relations (3).

Thus, we focused on three datasets. For the cross-paragraph relations (dataset 1), we analyzed explicit discourse relations which connect the first sentence of the paragraph to any previous text. As for local relations (2), the set in question concerns explicit inter-sentential discourse relations that do not go beyond the scope of one paragraph. The relations in the set (3) connect arguments within one syntactic tree (sentence), i.e. they cross neither the sentence boundary, nor the paragraph boundary.<sup>22</sup>

The distribution of semantic types of discourse relations across all the datasets differs significantly, see Table 6 and the graph in Figure 2, the significance was verified using the  $\chi^2$  test.

The vast majority of instances (63-68 percentage points in all three datasets) belong to three discourse relations, namely *conjunction*, *opposition*, and *reason-result*. In cross-paragraph relations, *conjunction* comes first, whereas in intra-paragraph inter-sentential relations, *opposition* is most common. Within intra-sentential relations, *conjunction* absolutely predominates with 42% of occurrences, and the very frequent relation of *condition* comes as third, compare details below.

In the top ten of the cross-paragraph relations, some relations typically occur which are not in the top ten of the intra-paragraph relations, namely *specification*, *generalization*, and *instantiation*. Often, the following paragraphs in our data expand the content of the previous text in this way. On the other hand, we cannot consider the relation of *specification* as only typical in the higher

22. Theoretically, there could be cross-paragraph intra-sentential discourse relations, too, such as lists of dependent clauses printed each in a different paragraph, cf.

*The request will be accepted*  
- **if** the applicant comes from the EU, **and**  
- **if** he/she encloses two letters of recommendation.

We do not deal with such sentences in this paper since their occurrence is marginal.

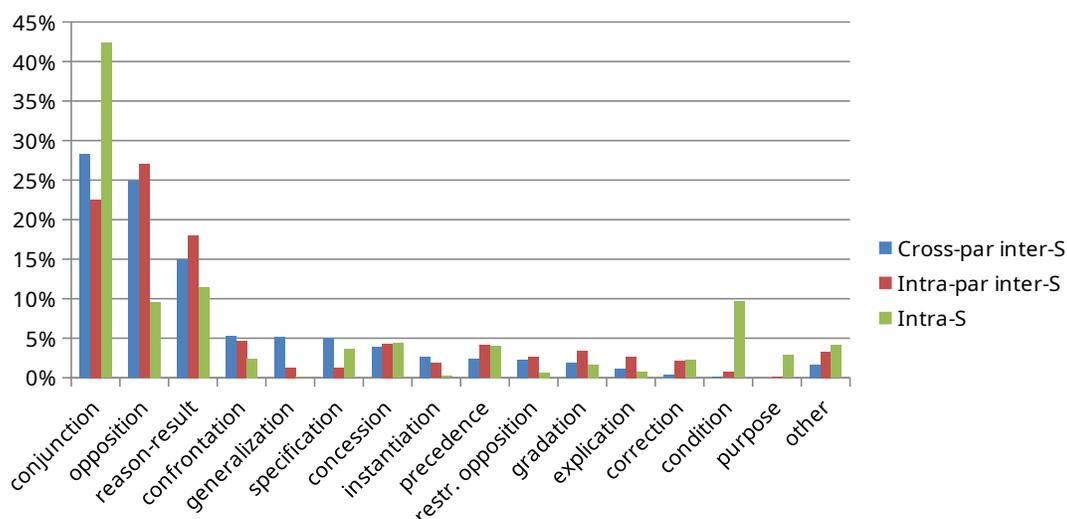


Figure 2: A graph for Table 6 showing percentages of the semantic types for the three datasets, i.e. for cross-paragraph (first column), intra-paragraph inter-sentential (second column), and intra-sentential discourse relations (third column). In each dataset, all its occurrences sum to 100%.

text structure, since it also occurs in the top 10 intra-sentential relations. We can assume that some structural features of *specification* in the datasets (1) and (3) differ; these features will be an object to a further research. In the top ten of the intra-paragraph inter-sentential relations (dataset 2), on the other hand, *gradation*, *explication* and *correction* are present, representing probably another way of text progression typical for smaller units.

The datasets (1-3) differ not only in proportions of frequent semantic types, but also in those which are **the least** represented in the respective datasets. We have looked into semantic types of discourse relations which are represented by less than 1% of occurrences in each of the analyzed datasets, see Table 7.

As can be seen from Table 7, some relations are rare in our data in general, independently from paragraph boundaries and syntactic structure. This is the case of all the *pragmatic relations*, *equivalence* and *conjunctive alternative*. Another group of relations has a low representation in the dataset (1) and (2), but they are quite typical for intra-sentential relations (dataset 3). This applies to *condition* and *purpose*, see Table 6 and Table 7. We can consider these relations as typically syntactic and local. The relation of *correction* is usually not used across paragraphs. However, it is typical for both types of intra-paragraph relations. Finally, we can find typical relations of the higher structure in our data (dataset 1), too, which occur in the intra-sentential structure (dataset 3), namely *instantiation* and *generalization*.

Let us sum up the observations concerning semantics of discourse relations in the datasets (1-3). Distributions of semantic types in inter-sentential relations (datasets 1 and 2) are quite close to each other, but they differ from intra-sentential relations distinctly. This finding was confirmed by the result of  $\chi^2$  tests, too. In other words, inter/intra-sententiality is a more important feature for the semantics of discourse relations than the presence or absence of paragraph boundary. Nevertheless,

(1) Cross-par inter-S		(2) Intra-par inter-S		(3) Intra-S	
equivalence	8	equivalence	57	explication	115
synchrony	3	synchrony	49	restr. opposition	99
correction	3	condition	42	conj. alternative	70
pragm. reason-result	2	pragm. contrast	27	equivalence	42
condition	1	pragm. reason-result	27	instantiation	27
disj. alternative	1	conj. alternative	20	pragm. contrast	23
conj. alternative	0	disj. alternative	14	pragm. reason-result	16
purpose	0	purpose	7	pragm. condition	16
pragm. contrast	0	pragm. condition	1	generalization	12
pragm. condition	0				

Table 7: Distribution of semantic types with occurrence <1% for each dataset (number of occurrences)

cross-paragraph relations (dataset 1) have specific features distinguishing them from intra-paragraph relations (datasets 2 and 3), too. Besides the most frequent relations the distributions of which are very similar for all the three datasets (*conjunction*, *opposition*, *reason-result*), cross-paragraph relations typically express different meanings of expansion (*generalization*, *specification*, *instantiation*). On the other hand, they rarely express semantics of *correction* which is typical for intra-paragraph relations (datasets 2 and 3); moreover, some typical intra-sentential relations, such as *condition*, *purpose*, *disjunctive alternative* and *synchrony* almost do not occur across paragraphs in our data.

#### 4.3.1 DISCOURSE CONNECTIVES IN PARAGRAPH-INITIAL POSITIONS

We also addressed the question of the differences in the use of discourse connectives across paragraphs and in local relations. We focused on the three most common relations, *conjunction*, *opposition* and *reason-result*, and the connectives by which these relations are expressed (see Table 8). The analysis showed that the relation of *conjunction* is often expressed by specific discourse connectives in the cross-paragraph context, namely with the adverb *dále* [*further*, *next*] and the particle *také* [*too*, *also*]. The basic and most frequent discourse connective for *conjunction* is *a* [*and*] is much more frequently used in the datasets (2) and (3), i.e. in the intra-paragraph structure. Further, in the intra-paragraph coherence, discourse connectives based etymologically on a prepositional phrase with an anaphoric element are common, cf. *přítom* [*at the same time*, lit. *by that*], and its intra-sentential relative variant *přičemž* [*and/while*, lit. *by which*]. Discourse connectives based on relatives are not used in the inter-sentential structures.

Discourse connectives for *opposition* are almost identical for cross-paragraph and intra-paragraph inter-sentential relations (dataset 1 and 2). A typical discourse connective in both these groups is *však* [*however*], whereas intra-sententially (dataset 3), *opposition* is predominantly expressed by the conjunction *ale* [*but*]. In dataset (3), multi-part discourse connectives are used (with a non-autonomous part *sice* [*sure/true*]) which are not used inter-sententially.

Similarly, discourse connectives used for *reason-result* are very close in the datasets (1) and (2). Again, a specific set of discourse connectives is used in the intra-sentential context (dataset 3). Moreover, these two larger groups (inter-sentential datasets 1 and 2, and the intra-sentential set 3)

(1) Cross-par inter-S	(2) Intra-par inter-S	(3) Intra-S
5 most common discourse connectives: conjunction		
dále [further]	42 a [and]	309 a [and] 5 364
také [too]	32 také [too]	193 což [which] 165
a [and]	28 dodal [he added]	78 přičemž [by which] 71
rovněž [also]	25 rovněž [also]	74 a také [and also] 45
dodal [he added]	11 přitom [at the same time]	73 podobně [similarly] 39
5 most common discourse connectives: opposition		
však [however]	123 však [however]	906 ale [but] 732
ale [but]	33 ale [but]	310 však [however] 221
ovšem [yet]	24 ovšem [yet]	154 sice - ale [true/sure - but] 143
avšak [however]	4 jenže [but/only]	39 ovšem [yet] 45
jenže [but/only]	4 přitom [at the same time]	31 sice - však [although - yet] 37
5 most common discourse connectives: reason-result		
proto [therefore]	31 proto [therefore]	301 protože [because] 510
tedy [thus]	25 totiž [namely/you see]	285 neboť [for] 212
totiž [namely/you see]	21 tedy [thus]	142 takže [so] 116
a tak [and so]	6 tak [so]	66 a tak [and so] 95
v tomto směru [in this sense]	5 a tak [and so]	34 proto, že [because] 95

Table 8: The most common discourse connectives for conjunction, opposition, and reason-result (number of occurrences)

mostly differ in the direction of the *reason-result* relation which influences the choice of the discourse connectives. While inter-sentential relations mainly express the meaning of result (discourse connectives *proto [therefore]*, *tedy [thus]*, *a tak [and so]*, *v tomto směru [in this sense]*), intra-sentential relations more typically represent the meaning of reason (discourse connectives *protože [because]*, *neboť [for]*, *proto, že [because]*).

Generally, there is a certain difference among the observed datasets in the usage of secondary connectives. Whereas in intra-sentential relations (dataset 3) secondary connectives do not occur in high positions of the table, they are yet frequent enough in the dataset 2 (*dodal [he added]*) and even more in the dataset 1 (*dodal [he added]*, *v tomto směru [in this sense]*).

To sum up, discourse connectives in the cross-paragraph and local coherence overlap to a large extent; nevertheless, paragraph-initial discourse connectives still have some special features. Some of them are based on the inter-sentential character of cross-paragraph relations and they are common for datasets (1) and (2). Thus, discourse connectives in these groups do not include relative elements or typical intra-sentential expressions (*sice [true/sure]*). Further, in the inter-sentential context, discourse connectives expressing result are used more often than connectives of reason. The latter are, on the other hand, more frequent in the intra-sentential relations (dataset 3).

Within the group of inter-sentential discourse connectives (datasets 1 and 2), paragraph-initial connectives differ from intra-paragraph connectives in certain aspects. For the relation of *conjunction*, the specific discourse connective *dále [further, next]* is frequently used. The proportion of secondary discourse connectives is higher in this group than in datasets (2 and 3), too.

Left Arg size (no. of sentences)	Right Arg size (no. of sentences)	Count	Percent
1	1	5 966	89.1%
1	large	92	1.4%
large	1	617	9.2%
large	large	20	0.3%
total		6 695	100%

Table 9: Overview of argument sizes of inter-sentential discourse relations without lists in PDiT 2.0.

#### 4.4 Large arguments

Although the *minimality principle* was taken into account during annotating argument extents in the PDiT, the annotators could also mark large argument spans and non-adjacent arguments, if justified (compare 4.2). In a way, the minimality principle, if applied thoroughly, can prevent the detection of natural “whole paragraph” to “whole paragraph” relations and similar. In this part of the analysis, we try to look into cases where marking a large argument was superior to the minimality requirement, we quantify them and try to find explanations for them.

For this study, relations with large arguments are specified as relations where either one or both arguments are larger than one sentence. For the analysis, we only take into account explicit inter-sentential relations and exclude list structures. We also exclude such intra-sentential relations, which cross the sentence boundary with a part of one of their arguments. They represent a very specific group (25 instances in the corpus).

In the data of the PDiT 2.0, extents of discourse arguments are given by the tree representation of a sentence – a discourse relation is marked between two tree nodes, roots of two subtrees that in most cases represent the arguments – and further specified by two range attributes, `start_range` and `target_range`, which help define more complex cases. It is important to keep in mind that for symmetric relations (i.e. where both arguments are of the same nature and the linear order of the arguments is always the same, as in *conjunction*, *opposition* or *synchrony*), the `target_range` attribute value always defines the extent of the left-sided argument and the `start_range` attribute value defines the extent of the right-sided argument. For asymmetric relations like *reason-result*, in which the arguments can switch the order, `start_range` and `target_range` are assigned to each relation individually via semantic definitions of Arg1 and Arg2. Arg1 always has the `target_range` attribute, disregarding its location. This property of the PDiT annotation needs to be taken into account in the phase of the analysis below where the left or right position of the arguments matter.<sup>23</sup>

The proportions of inter-sentential relations with different sizes of arguments are summarized in Table 9. There are 5 966 relations with single-sentenced both arguments in the PDiT 2.0, in other words relations with “small” arguments, and they represent 89% of all inter-sentential relations.

23. The `target` arguments (Arg1) are the following for asymmetric relations: (pragm.) result for the relation(s) of (pragm.) *reason-result*, statement being explained (*explication*), result of a (pragm.) condition ((pragm.) *condition*), denial of expectation (*concession*), statement being corrected (*correction*), statement having a purpose (*purpose*), lower degree (*gradation*), successive event (*precedence-succession*), statement being restricted/having an exception (*restr. opposition*), more general statement (*instantiation, specification*), more specific statement (*generalization*). For symmetric relations, target argument (Arg1) is always the left-sided argument.

Relation	Total	Left Arg - Right Arg			Start (Arg2) - Target (Arg1)		
		Large Left	Large Right	Both Large	Large Start	Large Target	Both Large
concession	29	27	1	1	23	5	1
condition	3	3	0	0	2	1	0
confrontation	39	35	2	2	2	35	2
conjunction	115	107	4	4	4	107	4
conj. alternative	2	1	1	0	1	1	0
correction	6	5	0	1	0	5	1
equivalence	11	8	0	3	0	8	3
instantiation	21	6	13	2	13	6	2
explication	23	7	14	2	15	6	2
generalization	78	76	2	0	2	76	0
gradation	13	10	2	1	2	10	1
opposition	173	158	12	3	12	158	3
pragm. contrast	3	2	1	0	1	2	0
pragm. reason-result	6	6	0	0	3	3	0
precedence-succession	9	9	0	0	9	0	0
reason-result	176	139	36	1	155	20	1
restr. opposition	13	12	1	0	1	12	0
specification	3	0	3	0	3	0	0
synchrony	6	6	0	0	0	6	0
total	729	617	92	20	248	461	20

Table 10: Large arguments in PDiT 2.0, measured both for the left-right positions and start-range target-range directions (= semantics of the arguments); "large" means that the given argument spans more than one sentence while the other one (in the same column) only spans one sentence (of its subset); symmetric relations are marked with the light gray background.

The remaining 729 relations that have at least one large argument (two and more sentences), form 11% of all inter-sentential relations. There are 617 relations with a large left argument and a single-sentenced right argument (large – 1), compared to only 92 relations with large right argument and a single-sentenced left argument (1 – large). There are only 20 relations with both large arguments (large – large), which is, given the figures for hierarchies, in our opinion a surprisingly low number, in terms of percentage negligible.

The most distinctive finding for large arguments is the huge disproportion of left-sided and right-sided large arguments, or in other words, the great predominance of large left-sided arguments. In a detailed view, the figures in Table 10, the “left - right” section, show that this is the case across almost all discourse types (senses).<sup>24</sup> However, the left-right positions of the arguments are only informative for symmetric relations (with grey background in Table 10), we will elaborate on

24. In our data, three relations, *disjunctive alternative*, *pragm. condition* and *purpose* do not have any instances with large arguments.

them further in this section.<sup>25</sup> For asymmetric relations, argument semantics plays a crucial role. There are three exceptions to the tendency of a larger left-sided argument in Table 10: *instantiation*, *explication* and *specification*, where large right-sided arguments are more common. But, if we look at the “start-target” part of the table, which is more informative for these asymmetric relations, we can see that in all these cases, argument semantics goes hand in hand: these are precisely the relations where the right contexts are represented by those arguments, which are easily conceivable as more elaborated, expandable: the argument providing an explanation (large start arg = 15), the argument giving example (13), the specifying argument (3).<sup>26</sup> For comparison, an analogous, and in terms of numbers much stronger, disproportion is visible in *generalization*, also an asymmetric relation, with 76 large left contexts which are also the 76 more specific arguments. The figures here fully comply with the intuitive notions of how semantics of these relations work, e.g. a generalizing, summarizing statement should be generalizing over a large previous text segment.

What could be less intuitive are the sizes of individual arguments in *reason-result* relation. For this relatively frequent relation,<sup>27</sup> (176 instances with large arguments in total), 155 (88%) have a large “reason” arguments but only 20 (11%) have a large “result” argument. The large arguments stand mostly on the left. Thus, semantically, the arguments are in line with the similar *explication* relation, that means the dominance of large reason/explication arguments, but in terms of location, these arguments stand elsewhere, on the left for *reason-result* and on the right for *explication*. A possible explanation can be the different importance of the arguments, in the terms of RST nuclearity, and/or role of secondary connectives and connective phrases.

For symmetric relations, we do not have a straightforward explanation for the predominance of large left-sided arguments. Only the following assumptions can be suggested: we might evaluate this as an effect of the annotation strategy, given that the connective is mostly a part of the right-sided argument, and so it may seem unnatural to go beyond the strong right boundary of the connective-containing sentence. We may also ask if the minimality principle is applied the same way to left and right contexts. And/or, this phenomenon might be inherent to language. In a similar manner in which anaphora in the language occurs much more often than cataphora, the connectives (some of which are indeed anaphoric, compare Webber et al. (2003), Stede and Grishina (2016) or Poláková and Mírovský (2019)), relate to small or larger previous semantic contents.<sup>28</sup> From a cognitive perspective, this disproportion of argument sizes may be connected with the linear way of text production and also the gradual growth of information received by the reader, and the perspective of the annotator, who may proceed incrementally, like a reader, not knowing about the sizes of any right context. This issue needs a further insight, since it may be very important for the understanding of the difference of analytic perspectives in local and global annotation approaches.

25. This also implies that proportions of numbers for symmetric relations in the left-right section and the start-target section of Table 10 are identical, they only reflect the annotation convention that the target argument is always on the left, in other words, the discourse arrow always leads to the left for the symmetric relations.

26. Counts for these three relations are too low to draw any hard conclusion but even the small numbers here support the intuitive claims.

27. Even when its intra-sentential instances are filtered out here, there are no *protože* [*because*]-relations included.

28. In our experience, cataphoric connectives are mostly secondary, with a demonstrative element that mostly introduces a dependent clause (e. g. *thanks to the fact that...* ) and so their scope is very narrow. This might be, however, language-specific.

## 5. Conclusion

The aim of this study was to determine, using corpus methods, to what extent local annotations of discourse relations enable to abstract and describe phenomena of global coherence or higher text structuring. We used the 50 thousand sentences of the Prague Discourse Treebank 2.0 for Czech and the equally sized Penn Discourse Treebank 3 for English.<sup>29</sup> The analysis focused on three main aspects: 1. the “shape” of the text in terms of mutual configuration of close discourse relations (pairwise); 2. cross-paragraph relations, their semantics and the properties of connectives in these relations as opposed to intra-paragraph settings, and 3. the size of arguments (text units) connected by discourse relations.

Regarding 1., even the discourse relations annotated in local annotations settings are assumed to form specific patterns, which are inherent in global coherence models like the Rhetorical Structure Theory (RST), but not postulated for local coherence models. This includes recursive hierarchical structuring of smaller and larger relations. Also, the RST model applies strong constraints on the overall document structure, defining it as a (constituency) tree with no crossings and overlaps.

Our analysis of relation configurations further contributes to the theoretical coherence-oriented research in particular by bringing empirical data to the discussion whether RST tree graphs are adequate (and sufficient) to represent discourse structure: Czech data available to us support this claim, exceptions are very rare. We have described patterns that are typical (adjacency, progress, hierarchy, etc.), less typical (argument containment patterns, envelopment) and quite rare (total overlap, crossing, partial overlaps etc.) in our data, and analyzed them linguistically. Further, we have compared our findings to those of a similar study conducted on English PDTB, version 2 (Lee et al., 2006), learning that the proportions of occurrence of individual patterns roughly correspond in both corpora, although our study distinguishes some more subtle configurations. Frequent patterns in our data comply with the RST tree structure rules. Less frequent patterns in the PDiT mostly deal with inclusion or exclusion of attribution spans, but also with the annotation strategies for secondary connectives in cases where they form a whole clause (*It means that...*) or they are anaphoric (*in this respect*). In some rare patterns, where, in our opinion, there is a violation of the tree structure in the sense of RST, we have found a small number of linguistically defensible interpretations that are not to be factored out due to discourse anaphora or attribution, as Lee et al. (2006) suggest.

Next, we have investigated hierarchies built by nested local relations in both Czech and English data. In all investigated properties in this respect, the two corpora are very similar. In both of them, we have detected even 5-level hierarchies, although they are quite rare. In a more detailed perspective, however, much of the structure proved to be intra-sentential: beyond the sentence boundary, local annotation of explicit connectives and AltLexes does not expose hierarchical text structuring very often. Detected hierarchies of inter-sentential relations only reached max. 3 levels of depth in both corpora. We do not claim that the trees detected by us are the trees a global analysis like RST would discover, but we demonstrate the existence of some hierarchical text structure in local annotation.

2. In the part of the analysis concerning cross-paragraph phenomena, the distributions of semantic types of discourse relations reveal dominance of three elaborative meanings, namely *specification*, *generalization*, and *instantiation* in relations crossing the paragraph boundary. These relations

---

29. The analysis of English locally annotated data is so far only complementary to the analyses of Czech data but we plan to extend it also to other subtopics in this study in the future.

are not in the top ten most frequent of all other intra-paragraph relations,<sup>30</sup> and our findings only confirm their intuitively perceived large role in text composition and lesser role in syntax.

Nevertheless, it was observed that the feature of inter/intra-sententiality is more important for the semantics of discourse relations than the presence or absence of the paragraph boundary. Typical intra-sentential relations, such as *condition*, *purpose*, *disjunctive alternative* and *synchrony*, almost never occur in cross-paragraphs relations in our data.

As for connectives, for the relation of *conjunction*, there seems to exist a specific discourse connective of cross-paragraph links, the connective *dále [further; next]*. Distributions of other connectives in frequent relations between and within paragraphs do not vary much, and, counter-intuitively, also coordinating conjunctions (*and*, *but*) are fairly represented in cross-paragraph relations. The proportion of secondary discourse connectives is higher in these contexts.

3. The analysis of argument size examined the hypothesis that in local coherence models, existence of large arguments (more than one sentence) should be limited by the annotation principle to annotate minimal units. It was discovered that relations with one or both large arguments are indeed not very frequent in PDiT 2.0 (11% of all inter-sentential relations) and that the large argument is in almost 85% on the left, which might be annotator's bias when proceeding from left (known context) to right (unknown context), and/or it may be an inherent property of texts. It would be interesting to compare this observation to the ways of tree branching in the RST-Discourse Treebank global annotations.

In the future, we plan to further extend our analysis to the PDTB 3 data and we would like to include also implicit relations, entity-based relations and hypophora (relations of question and answer) in both languages. Implicitness is an important feature of inter-sentential discourse relations, therefore inclusion of implicit relations into the research will enhance the general picture of the distribution of semantic types. This kind of results can be used then e.g. for the prediction of meaning of inter-sentential discourse relations. Furthermore, the role of the typically implicit semantic types, such as instantiation or specification, can be described in detail in this way.

The outcome of the study will be reflected in a future RST-like annotation of Czech. First, the results will be confronted with a real pilot RST analysis on a sample of locally annotated Czech texts with detected hierarchical organization, in order to assess the degree of equivalence of the hierarchical structures. The findings about hierarchies and the large arguments from this study can be further crosschecked with notions like nuclearity and canonical order of discourse arguments to find more about possible bridging through global and local frameworks. Also, the distribution of semantic labels given the size of text arguments in local annotations can be related to the use of RST labels (e.g. to the division to subject matter and presentational rhetorical relations) and their correspondence in lower and higher structuring can be discussed.

The fact that local discourse annotation in both PDiT and PDTB also displays hierarchical structure (up to 5 levels of depth) but at least two lowest levels are usually intra-sentential, implies a large role of syntax in discourse complexity. Syntactic hypotaxis/parataxis, but also the (a)symmetry of local discourse labels when related to nuclearity can be of advantage in a possible automatic RST pre-annotation or in rhetorical parsing.

Last but not least, methodologically, our experiments seem to reveal a lot about annotation strategies and biases: the minimality principle seems to affect the left-sided and right-sided argument sizes with great difference and its consistent application also may hinder the ability of local

---

30. with the sole exception of *specification* in intra-sentential use, which may be connected to the very frequent nominal right-sided arguments (and governing verb ellipsis) in the PDiT annotation.

models to accurately assess coherence of larger blocks. Our results also open space for the hypothesis that the annotation procedure itself, e.g. the order in which individual segments are connected to other segments, may influence the segment size and the hierarchical structure formed.

## Acknowledgements

The authors would like to thank three anonymous reviewers for their detailed and constructive comments. The authors acknowledge support from the and the Grant Agency of the Czech Republic (project no. 20-09853S). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2018101).

## References

- Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.
- John A. Bateman and Klaas Jan Rondhuis. Coherence Relations: Towards a General Specification. *Discourse processes*, 24(1):3–49, 1997. URL <https://www.tandfonline.com/doi/abs/10.1080/01638539709545006>.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer, 2003. URL [https://link.springer.com/chapter/10.1007/978-94-010-0019-2\\_5](https://link.springer.com/chapter/10.1007/978-94-010-0019-2_5).
- Iria Da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. On the development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10. Association for Computational Linguistics, 2011.
- Debopam Das, Maite Taboada, and Paul McFetridge. *RST Signalling Corpus*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, 2015. URL <https://catalog.ldc.upenn.edu/LDC2015T10>.
- Markus Egg and Gisela Redeker. How Complex is Discourse Structure? In *Proceedings of LREC 2010, Malta*, 2010. URL [https://lexitron.nectec.or.th/public/LREC-2010\\_Malta/pdf/796\\_Paper.pdf](https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/796_Paper.pdf).
- Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949, 2014. URL <https://www.aclweb.org/anthology/C14-1089.pdf>.
- Barbara Grosz and Candace Sidner. Attentions, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986. URL <https://dash.harvard.edu/handle/1/2579648>.
- Caroline Hagege and Guillaume Jacquet. Combining Temporal Processing and Textual Entailment to Detect Temporally Anchored Events, December 18 2014. US Patent App. 13/920,462.

- Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Uřešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. *Prague Dependency Treebank 3.5*. Data/Software, Univerzita Karlova, MFF, ÚFAL, Prague, Czech Republic, 2018. URL <http://hdl.handle.net/11234/1-2621>.
- Jerry R. Hobbs. Coherence and Coreference. *Cognitive science*, 3(1):67–90, 1979. URL [https://doi.org/10.1207/s15516709cog0301\\_4](https://doi.org/10.1207/s15516709cog0301_4).
- Mikel Iruskieta, Maria J Aranzabe, Arantza Diaz de Ilarraza, Mikel Lersundi, and Oier Lopez de Lacalle. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop RST and Discourse Studies*, pages 40–49, 2013.
- Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, 2016.
- Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex than in Syntax? In *Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories*, pages 12–23, 2006. URL <http://ufal.mff.cuni.cz/t1t2006/pdf/133.pdf>.
- Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse III Workshop*, pages 61–68, 2008.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. Automatically Evaluating Text Coherence Using Discourse Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, 2011. URL <https://www.aclweb.org/anthology/P11-1100.pdf>.
- Haitao Liu and Hongxin Zhang. Rhetorical relations revisited across distinct levels of discourse unit granularity. *Discourse Studies*, 18(4):454–472, 2016. doi: 10.1177/1461445616647891.
- William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8:243–281, 1988. URL <https://semanticsarchive.net/Archive/GMyNDBjO/RST%20towards%20a%20functional%20theory%20of%20text%20organization.pdf>.
- Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT press, 2000.
- Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Annotating discourse connectives and their arguments. *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, 2004.

- Jiří Mírovský, Pavlína Jínová, and Lucie Poláková. Does Tectogrammatcs Help the Annotation of Discourse? In *Proceedings of COLING 2012: Posters*, pages 853–862, 2012. URL <https://www.aclweb.org/anthology/C12-2083.pdf>.
- Jiří Mírovský, Pavlína Synková, and Lucie Poláková. Extending Coverage of a Lexicon of Discourse Connectives. *Prague Bulletin of Mathematical Linguistics*, in prep.
- Petr Pajas and Jan Štěpánek. Recent Advances in a Feature-rich Framework for Treebank Annotation. In Donia Scott and Hans Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester, 2008. The Coling 2008 Organizing Committee. URL <https://www.aclweb.org/anthology/C08-1085.pdf>.
- Petr Pajas and Jan Štěpánek. System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, 2009. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P09-4009.pdf>.
- Lucie Poláková and Jiří Mírovský. Anaphoric Connectives and Long-Distance Discourse Relations in Czech. *Computación y Sistemas*, 23(3):711–717, 2019. URL <https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3274>.
- Lucie Poláková and Jiří Mírovský. Mining Local Discourse Annotation for Features of Global Discourse Structure. In *Lecture Notes in Artificial Intelligence, 23rd International Conference on Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 50–60, Cham, Switzerland, 2020. Faculty of Informatics, Masaryk University Brno, Springer. ISBN 978-3-030-58322-4. URL [https://link.springer.com/chapter/10.1007/978-3-030-58323-1\\_5](https://link.springer.com/chapter/10.1007/978-3-030-58323-1_5).
- Lucie Poláková, Jiří Mírovský, and Pavlína Synková. Signalling Implicit Relations: A PDTB-RST Comparison. *Dialogue & Discourse*, 8(2):225–248, 2017. URL <https://firstmonday.org/ojs/index.php/dad/article/view/10787>.
- Aurora Pons-Porrata, Rafael Berlanga-Llavori, and José Ruiz-Shulcloper. Topic Discovery Based on Text Mining Techniques. *Information processing & management*, 43(3):752–768, 2007.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, 2008. European Language Resources Association. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.9566&rep=rep1&type=pdf>.
- Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Exploiting Scope for Shallow Discourse Parsing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL <https://www.aclweb.org/anthology/L10-1634/>.
- Rashmi Prasad, Katherine Forbes Riley, and Alan Lee. Towards Full Text Shallow Discourse Relation Annotation: Experiments with Cross-Paragraph Implicit Relations in the PDTB. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 7–16, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5502. URL <https://www.aclweb.org/anthology/W17-5502>.

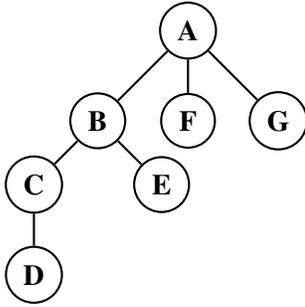
- Rashmi Prasad, Bonnie Webber, and Alan Lee. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, 2018.
- Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. *Penn Discourse Treebank Version 3.0*. Data/Software, Linguistic Data Consortium, 2019. URL <https://catalog.ldc.upenn.edu/LDC2019T05>. University of Pennsylvania, Philadelphia. LDC2019T05.
- Magdaléna Rysová and Kateřina Rysová. The Centre and Periphery of Discourse Connectives. In *Proceedings of Pacific Asia Conference on Language, Information and Computing*, pages 452–459, Bangkok, 2014. Department of Linguistics, Faculty of Arts, Chulalongkorn University. URL <https://www.aclweb.org/anthology/Y14-1052.pdf>.
- Magdaléna Rysová, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, and Šárka Zikánová. Prague Discourse Treebank 2.0. Data/Software. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2016. URL <http://hdl.handle.net/11234/1-1905>.
- Jerry Samet and Roger Schank. Coherence and Connectivity. *Linguistics and Philosophy*, 7(1): 57–82, 1984. URL <https://link.springer.com/article/10.1007/BF00627475>.
- Manfred Stede and Yulia Grishina. Anaphoricity in Connectives: A case study on German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 41–46, 2016.
- Manfred Stede and Arne Neumann. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of LREC 2014*, pages 925–929, Reykjavik, Iceland, 2014. URL [http://www.lrec-conf.org/proceedings/lrec2014/pdf/579\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/579_Paper.pdf).
- Maite Taboada and William C Mann. Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse studies*, 8(3):423–459, 2006. URL <https://journals.sagepub.com/doi/abs/10.1177/1461445606061881>.
- Christoph Unger. *Genre, Relevance and Global Coherence: The Pragmatics of Discourse Type*. Springer, 2006.
- Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545–587, 2003.
- Sandra Williams and Ehud Reiter. A corpus analysis of discourse relations for Natural Language Generation. In *Proceedings of Corpus Linguistics, 2003, Lancaster University, United Kingdom*, 2003. URL <http://ucrel.lancs.ac.uk/publications/CL2003/papers/williams.pdf>.
- Florian Wolf and Edward Gibson. Representing Discourse Coherence: A Corpus-Based Study. *Computational linguistics*, 31(2):249–287, 2005. URL <https://www.mitpressjournals.org/doi/abs/10.1162/0891201054223977>.

- Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. Discourse Graphbank, LDC2005T08 [Corpus]. *Philadelphia: Linguistic Data Consortium*, 2005. URL <https://catalog.ldc.upenn.edu/LDC2005T08>.
- Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. TED Multilingual Discourse Bank (TED-MDB): A Parallel Corpus Annotated in the PDTB Style. *Language Resources and Evaluation*, (54):587–613, 2019. URL <https://link.springer.com/article/10.1007%2Fs10579-019-09445-9>.
- Renxian Zhang. Sentence Ordering Driven by Local and Global Coherence for Summary Generation. In *Proceedings of the ACL 2011 Student Session*, pages 6–11, 2011.
- Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia, 2015.
- Šárka Zikánová, Jiří Mírovský, and Pavlína Synková. Explicit and Implicit Discourse Relations in the Prague Discourse Treebank. In *Proceedings of the 22nd International Conference on Text, Speech and Dialogue - TSD 2019*, volume 11697 of *Lecture Notes in Computer Science*, pages 236–248, Cham / Heidelberg / New York / Dordrecht / London, 2019. University of West Bohemia, Springer International Publishing. ISBN 978-3-030-27946-2. URL [https://link.springer.com/chapter/10.1007/978-3-030-27947-9\\_20](https://link.springer.com/chapter/10.1007/978-3-030-27947-9_20).

### Appendix 1: A four-level hierarchy in Prague Discourse Treebank 2.0

A Czech original and an English translation of a PDiT 2.0 text segment (cmpr9410\_034, 1.5 paragraphs, 11 sentences) with a detected 4-level hierarchy of discourse relations.

The pattern of the hierarchy is A ( B ( C ( D ) E ) F G ).



Context:

(1) Přehlídka bezradnosti [paragraph boundary] (2) Z historie je znám ne jeden případ, kdy na mistrně vyvedeném padělků pohorela celá plejáda soudních znalců. (3) Ve 30. letech tohoto století plnila titulní stránky berlínských denků jedna z největších afér v historii falzifikátů. (4) Obchodník s obrazy, jakýsi Wacker, byl obviněn, že dodal na výstavu 33 falešných van Goghů. (5) Skandál brzy přerostl v odbornou polemiku. (6) Obžalovací spis obsahoval několik tisíc stran.

(7) Soud vzal na vědomí názory více než šedesáti uměleckých historiků a kritiků.

(8) Nebyly mezi nimi ani dva, které by se nerozcházel alespoň v podrobnostech.

(9) Každé stanovisko našlo svého zastánce.

(10) Všichni začali být tou přemírou odborníků unaveni.

[paragraph boundary]

(11) Umělecký restaurátor a badatel de Wilt vzbudil v uměleckém světě senzaci, když (D) využil rentgenové záření k posouzení stáří obrazu.

(12) Jeho závěr zněl, že (C) soubor je starý minimálně tři desetiletí a tudíž pravý.

(13) Neboť (E) jaký blázen by měl zájem kolem roku 1895 padělat holandského umělce, jehož plátna byla k dostání po dvou stech francích?

(14) Obžaloba si těž (B) najala znalce, který pracoval s rentgenem.

(15) Ten prozkoumal kolekci z hlediska malířského rukopisu a (F) došel k závěru, že (G) ji nena-maloval van Gogh.

(16) Soud byl opět tam, kde předtím.

(17) Ani (A) znalci, kteří měli analyzovat obrazy chemickou cestou, nenašli společnou řeč.

Context:

(1) Parade of cluelessness [paragraph boundary] (2) There is more than one case in history in which a wide range of forensic experts have failed in judging a masterfully executed forgery. (3) In the 30's of this century, one of the greatest affairs in the history of falsehoods filled the front pages of Berlin newspapers. (4) A painting dealer, a Wacker, was accused of supplying 33 fake van Goghs to an exhibition. (5) The scandal soon became a professional controversy. (6) The indictment file contained several thousand pages.

(7) The court has recorded the views of over 60 art historians and critics.

- (8) There were not two of them that did not diverge at least in detail.  
 (9) Each opinion found its defender.  
 (10) Everyone grew weary of the excess of experts.  
 [paragraph boundary]  
 (11) The artistic restorer and explorer de Wilt caused a sensation in the world of art when (D) he used X-rays to assess the age of the painting.  
 (12) His conclusion was that (C) the collection was at least three decades old and therefore genuine.  
 (13) For (E) what fool would be interested, circa 1895, in forging a Dutch artist whose canvases were available at two hundred francs each?  
 (14) The prosecution also (B) hired an expert to work with the X-ray device.  
 (15) He examined the collection in terms of the painting's handwriting and (F) concluded that (G) it was not painted by van Gogh.  
 (16) The Court was back where it had been before.  
 (17) Not even (A) the experts who were to analyze the paintings chemically could find common ground.

**Relation A ('ani' [not even], conjunction):**

**leftarg:** (7 - 16) The court has recorded the views of over 60 art historians and critics. There were not two of them that did not diverge at least in detail. Each opinion found its defender. Everyone grew weary of the excess of experts.

The artistic restorer and explorer de Wilt caused a sensation in the world of art when he used X-rays to assess the age of the painting. His conclusion was that the collection was at least three decades old and therefore genuine. For what fool would be interested, circa 1895, in forging a Dutch artist whose canvases were available at two hundred francs each? The prosecution also hired an expert to work with the X-ray device. He examined the collection in terms of the painting's handwriting and concluded that it was not painted by van Gogh. The Court was back where it had been before.

[Soud vzal na vědomí názory více než šedesáti uměleckých historiků a kritiků. Nebyly mezi nimi ani dva, které by se nerozcházely alespoň v podrobnostech. Každé stanovisko našlo svého zastávce. Všichni začali být tou přemírou odborníků unaveni.]

Umělecký restaurátor a badatel de Wilt vzbudil v uměleckém světě senzaci, když využil rentgenové zařízení k posouzení stáří obrazu. Jeho závěr zněl, že soubor je starý minimálně tři desetiletí a tudíž pravý. Neboť jaký blázen by měl zájem kolem roku 1895 padělat holandského umělce, jehož plátna byla k dostání po dvou stech francích? Obžaloba si též najala znalce, který pracoval s rentgenem. Ten prozkoumal kolekci z hlediska malířského rukopisu a došel k závěru, že ji nenamaloval van Gogh. Soud byl opět tam, kde předtím.]

**rightarg:** (17) the experts who were to analyze the paintings chemically could find common ground  
 [znalci, kteří měli analyzovat obrazy chemickou cestou, nenašli společnou řeč]

**Relation B ('též' [also], conjunction):**

**leftarg:** (11 - 13) The artistic restorer and explorer de Wilt caused a sensation in the world of art when he used X-rays to assess the age of the painting. His conclusion was that the collection was at least three decades old and therefore genuine. For what fool would be interested, circa 1895, in forging a Dutch artist whose canvases were available at two hundred francs each?

[Umělecký restaurátor a badatel de Wilt vzbudil v uměleckém světě senzaci, když využil rentgenové záření k posouzení stáří obrazu. Jeho závěr zněl, že soubor je starý minimálně tři desetiletí a tudíž pravý. Neboť jaký blázen by měl zájem je kolem roku 1895 padělat holandského umělce, jehož plátna byla k dostání po dvou stech francích?]

**rightarg:** (14) The prosecution hired an expert to work with the X-ray device.

[Obžaloba si najala znalce, který pracoval s rentgenem.]

**Relation C ('Jeho závěr zněl, že' [His conclusion was that], generalization):**

**leftarg:** (11) The artistic restorer and explorer de Wilt caused a sensation in the world of art when he used X-rays to assess the age of the painting.

[Umělecký restaurátor a badatel de Wilt vzbudil v uměleckém světě senzaci, když využil rentgenové záření k posouzení stáří obrazu.]

**rightarg:** (sub12)<sup>31</sup> the collection was at least three decades old and therefore genuine

[soubor je starý minimálně tři desetiletí a tudíž pravý]

**Relation D ('když' [when], specification):**

**leftarg:** (sub11) The artistic restorer and explorer de Wilt caused a sensation in the world of art

[Umělecký restaurátor a badatel de Wilt vzbudil v uměleckém světě senzaci]

**rightarg:** (sub11) he used X-rays to assess the age of the painting

[využil rentgenové záření k posouzení stáří obrazu]

**Relation E ('neboť' [for], reason – result):**

**leftarg:** (sub12) (it is) genuine

[je pravý]<sup>32</sup>

**rightarg:** (13) what fool would be interested, circa 1895, in forging a Dutch artist whose canvases were available at two hundred francs each

[jaký blázen by měl zájem kolem roku 1895 padělat holandského umělce, jehož plátna byla k dostání po dvou stech francích]

**Relation F ('a' [and], conjunction):**

**leftarg:** (sub15) examined the collection in terms of the painting's handwriting

[prozkoumal kolekci z hlediska malířského rukopisu]

**rightarg:** (sub15) concluded that it was not painted by van Gogh

[došel k závěru že ji nenamaloval van Gogh]

**Relation G ('došel k závěru' [concluded that], generalization):**

**leftarg:** (sub15) He examined the collection in terms of the painting's handwriting

[prozkoumal kolekci z hlediska malířského rukopisu]

**rightarg:** (sub15) it was not painted by van Gogh

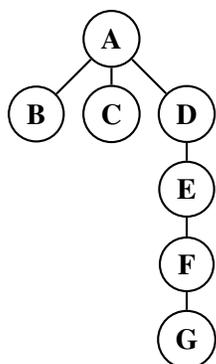
[ji nenamaloval van Gogh]

31. "sub" refers to the fact that the argument includes only a subset of a given sentence, not the whole sentence.

32. verb ellipsis restored from the previous clause via generated node in the tree structure

## Appendix 2: A five-level hierarchy in the Penn Discourse Treebank 3.0

The PDTB 3.0 text segment (wsj\_2431, 2 paragraphs, 7 sentences) with a detected 5-level hierarchy of discourse relations. The pattern of the hierarchy is **A ( B C D ( E ( F ( G ) ) ) ) )**.<sup>33</sup>



### Context:

(21) The coalition government tried to show that PASOK ministers had received hefty sums for OKing the purchase of F-16 Fighting Falcon and Mirage 2000 combat aircraft, produced by the U.S.-based General Dynamics Corp. and France's Avions Marcel Dassault, respectively. (22) Naturally, neither General Dynamics nor Dassault could be expected to hamper its prospective future dealings by making disclosures of sums paid (or not) to various Greek officials for services rendered.

(23) So it seems that Mr. Mitsotakis and his communist chums may have unwittingly served Mr. Papandreou a moral victory on a platter: PASOK, whether guilty or not, can now traipse the countryside condemning the whole affair as a witch hunt at Mr. Papandreou's expense.

(24) But (B) while (C) verbal high jinks alone won't help PASOK regain power, Mr. Papandreou should never be underestimated.

(25) First came his predictable fusillade: He charged the Coalition of the Left and Progress had sold out its leftist tenets by (F) collaborating in a right-wing plot aimed at ousting PASOK and (G) thwarting the course of socialism in Greece.

[paragraph boundary]

(26) Then (E), to buttress his credibility with the left, he enticed some smaller leftist parties to stand for election under the PASOK banner.

(27) Next (D), he continued to court the communists – many of whom feel betrayed by the left-right coalition's birth – by bringing into PASOK a well-respected Communist Party candidate.

(28) For balance, and in hopes of gaining some disaffected centrist votes, he managed to attract a former New Democracy Party representative and known political enemy of Mr. Mitsotakis.

(29) Thus (A) PASOK heads for the polls not only with diminished scandal-stench, but also with “seals of approval” from representatives of its harshest accusers.

33. Note that the relations B and C do not take part in the deepest branch of the hierarchy, they are only included in the higher relation A.

**Relation A ('thus', Contingency.Cause.Result):**

**leftarg:** (23 - 28) So it seems that Mr. Mitsotakis and his communist chums may have unwittingly served Mr. Papandreou a moral victory on a platter: PASOK, whether guilty or not, can now traipse the countryside condemning the whole affair as a witch hunt at Mr. Papandreou's expense. But while verbal high jinks alone won't help PASOK regain power, Mr. Papandreou should never be underestimated. First came his predictable fusillade: He charged the Coalition of the Left and Progress had sold out its leftist tenets by collaborating in a right-wing plot aimed at ousting PASOK and thwarting the course of socialism in Greece.

Then, to buttress his credibility with the left, he enticed some smaller leftist parties to stand for election under the PASOK banner. Next, he continued to court the communists – many of whom feel betrayed by the left-right coalition's birth – by bringing into PASOK a well-respected Communist Party candidate. For balance, and in hopes of gaining some disaffected centrist votes, he managed to attract a former New Democracy Party representative and known political enemy of Mr. Mitsotakis.

**rightarg:** (29) PASOK heads for the polls not only with diminished scandal-stench, but also with “seals of approval” from representatives of its harshest accusers

**Relation B ('but', Comparison.Concession.Arg2-as-denier):**

**leftarg:** (sub23) PASOK, whether guilty or not, can now traipse the countryside condemning the whole affair as a witch hunt at Mr. Papandreou's expense

**rightarg:** (24) while verbal high jinks alone won't help PASOK regain power, Mr. Papandreou should never be underestimated

**Relation C ('while', Temporal.Synchronous):**

**leftarg:** (sub24) But... Mr. Papandreou should never be underestimated

**rightarg:** (sub24) verbal high jinks alone won't help PASOK regain power

**Relation D ('next', Temporal.Asynchronous.Precedence):**

**leftarg:** (25 - 26) First came his predictable fusillade: He charged the Coalition of the Left and Progress had sold out its leftist tenets by collaborating in a right-wing plot aimed at ousting PASOK and thwarting the course of socialism in Greece.

Then, to buttress his credibility with the left, he enticed some smaller leftist parties to stand for election under the PASOK banner

**rightarg:** (sub27) he continued to court the communists ... by bringing into PASOK a well-respected Communist Party candidate

**Relation E ('then', Temporal.Asynchronous.Precedence):**

**leftarg:** (sub25) He charged the Coalition of of the Left and Progress had sold out its leftist tenets by collaborating in a right wing plot aimed at ousting PASOK and thwarting the course of socialism in Greece

**rightarg:** (sub26) he enticed some smaller leftist parties to stand for election under the PASOK banner

**Relation F ('by', Contingency.Cause.Reason):**

**leftarg:** (sub25) the Coalition of the Left and Progress had sold out its leftist tenets

**rightarg:** (sub25) collaborating in a right-wing plot aimed at ousting PASOK and thwarting the course of socialism in Greece

**Relation G ('and', Expansion.Conjunction):**

**leftarg:** (sub25) ousting PASOK

**rightarg:** (sub25) thwarting the course of socialism in Greece