Presenting the Prague Discourse Treebank 4.0

Jiří Mírovský, Pavlína Synková
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
mirovsky|synkova@ufal.mff.cuni.cz

Abstract

The Prague Discourse Treebank 4.0 is a large genre-diversified language resource with annotation of discourse relations marked by explicit connectives in Czech texts. It consists of 175 thousand sentences with 82 thousand discourse relations. We present the treebank as well as the methods used during the annotation of its individual parts, some of which were annotated fully manually, others using cost-effective partially automatic methods, achieving a comparable quality. The discourse annotation is available in two formats and theoretical frameworks: the Prague discourse annotation on top of deep syntax dependency trees, and the Penn Discourse Treebank style on top of plain texts, using both discourse type/sense taxonomies in both formats. The corpus is publicly and freely available, offering a valuable resource for linguistic research and natural language processing tasks.

Keywords: discourse relations, corpus, annotation, cost-efficient, discourse connectives

1. Introduction

Discourse analysis plays a pivotal role both in theoretical linguistic research and in natural language processing (NLP), enabling researchers and machines to interpret textual cohesion and coherence beyond individual sentences. At the core of discourse analysis is the identification and classification of discourse relations, which delineate semantic and pragmatic connections between clauses, sentences, or larger text segments. Discourse processing, which in a broader sense involves not only discourse relations, but also topic-focus articulation, coreference, bridging anaphora and other phenomena, supports a wide range of downstream applications including text summarization, machine translation, essay scoring, sentiment analysis, information extraction, question answering, and others (for an overview, see Joty et al., 2019).

The development of discourse-annotated corpora has been instrumental in advancing this field, providing essential data for linguistic studies and machine-learning methods. Discourse annotation efforts around the world go to various levels of detail, depth and complexity, following several theoretical approaches. The most influential examples are probably (i) the RST Discourse Treebank (Carlson et al., 2003) with a deep discourse analysis where – following the Rhetorical Structure Theory (Mann and Thompson, 1988) – each document forms a single annotation structure, and (ii) the Penn Discourse Treebank (Prasad et al., 2019), which has set a standard in shallow (or local) discourse annotation. Discourse-annotated resources for various languages either follow one (rarely even both) of the two mentioned approaches, with various degrees of modifications, or employ other theories entirely. From the most prominent ones, let us mention the Chinese Discourse TreeBank (Zhou and Xue, 2015), the Leeds Arabic Discourse Treebank (Al-Saif and Markert, 2010), the Hindi Discourse Relation Bank (Oza et al., 2009), the Turkish Discourse Bank (Zeyrek et al., 2010), the ANNODIS for French (Afantenos et al., 2012), and the Potsdam commentary corpus for German (Bourgonje and Stede, 2020), each employing distinct annotation schemes tailored to their needs. Recent contributions to discourse annotation are represented, for example, by corpora for Thai (Prasertsom et al., 2024) and Portuguese (Mendes and Lejeune, 2022); first steps to discourse annotation were done, e.g., for Japanese (Kubota et al., 2024) and Romanian (Mititelu and Voicu, 2024).

The Prague Discourse Treebank (PDiT) series is a long-term effort in creating such a resource for the Czech language. Since its first version, it has been heavily inspired by the shallow discourse annotation system of the Penn Discourse Treebank.

Initiated with PDiT 1.0 (Poláková et al., 2012), which annotated 49 thousand sentences of the Prague Dependency Treebank (Hajič et al., 2006) with explicit discourse relations, textual coreference and bridging anaphora (Poláková et al., 2013), the project has subsequently continuously evolved, leading to several published updates. PDiT 2.0 (Rysová et al., 2016) expanded the annotation to include secondary connectives,[1] while PDiT 3.0 (Synková et al., 2022) introduced a largely revised annotation of discourse rela-

[1] Primary connectives are fixed expressions such as *proto* [*therefore*] or *ale* [*but*], secondary connectives are less fixed (mostly multiword) expressions such as *z toho důvodu* [*for that reason*] or *jinými slovy* [*in other words*].

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6262–6276
11–16 May 2026.

tions, harmonized the annotation with the Lexicon of Czech Discourse Connectives (CzeDLex, Mírovský et al., 2017), and made the data available in both the native Prague Markup Language[2] and the Penn Discourse Treebank 3.0[3] format and taxonomy (Mírovský et al., 2023).

The present paper introduces the most recent release of the Prague Discourse Treebank, PDiT 4.0[4] (Synková et al., 2024), both in contrast with the previous versions, and as a stand-alone description of PDiT 4.0. The discourse annotation was significantly expanded in size to encompass all four subcorpora of the underlying Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0; (Hajič et al., 2024)): (i) the original Prague Dependency Treebank (PDT), (ii) the Czech part of the Prague Czech-English Dependency Treebank (PCEDT-cz), (iii) the Prague Dependency Treebank of Spoken Czech (PDTSC), and (iv) the Faust corpus. The new expansion of PDiT increases the total number of annotated sentences to over 175 thousand, with more than 82 thousand explicit discourse relations, making PDiT 4.0 one of the largest and most diverse discourse-annotated corpora available.

In Section 2, we introduce the annotation scheme of PDiT 4.0, both from the theoretical and technical points of view. Section 3 describes the annotation process of the individual parts of PDiT. Stronger emphasis is placed on the description of the PDTSC annotation, the details of which have not yet been published. We conclude in Section 4.

2. PDiT 4.0 Annotation Scheme

The term *discourse relation* refers to a semantic or pragmatic relation that connects two discourse units – segments of text expressing mostly individual events, states or situations (Zikánová et al., 2015; Zufferey and Degand, 2024). In Example 1, a discourse relation holds between two clauses and is signalled by an explicit discourse-structuring device, a connective *sice* ~ *ale* [*but*].[5]

(1) Sice *už mi bude 65*, ale **tancuju rád**. (PDTSC, pdtsc_088_1.01)
[*I may be turning 65*, but **I love to dance**.]

If a discourse relation is marked by a connective, we call it an *explicit* discourse relation. If the connective is absent, we call the relation *implicit*.

Each discourse relation in PDiT is assigned a single discourse type from a list of 22 Prague discourse types (in Example 1, *concession*) and – for compatibility with the Penn Discourse Treebank annotation scheme – a single sense from the Penn Discourse Treebank taxonomy (in Example 1, *Comparison.Concession.Arg2-as-denier*).

Table 1 shows the correspondence between the Prague discourse types and the two first levels in the taxonomy of Penn senses (the third level, e.g. *Arg2-as-denier*, represents the argument semantics, and it is indicated by the direction of the relation arrow in the PDiT approach, cf. Figure 1).

Details on the transformation method from the Prague discourse types to the Penn senses were given in Mírovský et al. (2023). The study was carried out on the data of PDiT 3.0 and led to the understanding that (i) approx. 42% of the discourse relations have a Prague discourse type that transforms unambiguously to a Penn sense (e.g., *synchrony* to *Temporal.Synchronous*), (ii) 56% of the relations can be reliably transformed using linguistically motivated rules (e.g., the ambiguity of *condition* between *Contingency.Condition* and *Contingency.Negative-condition* can be solved by the presence of negation and by the connectives), and (iii) only about 2% of the relations need to be manually disambiguated into a correct sense (e.g., for pragmatic relations, no formal cues could be found to distinguish SpeechAct from Belief (*SA* and *B* in Table 1)).

Inspired by the Penn Discourse Treebank annotation style, the Prague Discourse Treebank (i) follows the theory-neutral approach, i.e. it makes "no commitments to what kinds of high-level structures may be created from the low-level annotations of relations and their arguments" (Prasad et al., 2008), (ii) perceives discourse connectives as anchors for explicit relations,[6] and (iii) follows the minimality principle, i.e. the size of an argument, although it can encompass multiple clauses or sentences, is delimited only to the extent necessary for a correct interpretation of the relation.

2.1. PDiT Data Representation

In both theory and practice of the Prague approach, discourse relations are regarded as the

[2] PML (Hana and Štěpánek, 2012)
[3] PDTB (Prasad et al., 2019)
[4] https://ufal.mff.cuni.cz/pdit4.0
[5] We adopt here the Penn Discourse Treebank convention of highlighting two discourse arguments and the connective - Argument 1 (the left one in coordinated structures or in inter-sentential relations, or the governing one in subordinated structures) is typeset in italics, Argument 2 (the other argument) in bold and the connective is underlined.
[6] In the PDTB approach, implicit discourse relations are defined by their position (mostly between adjacent sentences or clauses in compound sentences if there is no connective between them). However, implicit discourse relations are not a part of the Prague Discourse Treebank 4.0 annotation. They were annotated only on a data sample (approx. 2.6 thousand sentences) and published as PDiT-EDA 1.0 (Zikánová et al., 2018).

PDiT discourse type	PDTB 3.0 sense(s)
COMPARISON	
concession	Comparison.Concession
confrontation	Comparison.Contrast
correction	Expansion.Substitution
gradation	Expansion.Conjunction
opposition	Comparison.Concession
pragm. contrast	Comparison.Concession+B, Comparison.Concession+SA, Comparison.Concession
restrictive opposition	Expansion.Exception, Comparison.Contrast
CONTINGENCY	
condition	Contingency.Condition, Contingency.Neg-condition
explication	Contingency.Cause+B, Expansion.Level-of-detail
purpose	Contingency.Purpose
pragm. reason–result	Contingency.Cause+B, Contingency.Cause+SA, Contingency.Cause,
pragm. condition	Contingency.Condition+SA, Contingency.Neg-condition+SA, Contingency.Condition
reason–result	Contingency.Cause, Contingency.Neg-cause
EXPANSION	
conjunction	Expansion.Conjunction, Comparison.Similarity
conj. alternative	Expansion.Disjunction
disj. alternative	Expansion.Disjunction
equivalence	Expansion.Equivalence
generalization	Expansion.Level-of-detail
instantiation	Expansion.Instantiation
specification	Expansion.Level-of-detail
TEMPORAL	
preced.–succession	Temporal.Asynchronous
synchrony	Temporal.Synchronous

Table 1: Basic transformation table from PDiT discourse types to the PDTB 3.0 second-level senses

superstructure of the syntactic analysis of a text, since the minimal units that the discourse relations connect are clauses and sentences, i.e. units that are clearly defined in syntactic trees and identified during syntactic analysis/parsing. In the Prague Discourse Treebank, annotation of discourse relations has been carried out on top of the deep-syntax layer (tectogrammatics, t-layer, Hajičová et al., 2008)[7] allowing the discourse structure to be linked not only to the syntactic structure, but also to all other levels of annotation in PDT.

[7] As was illustrated in several studies (Poláková et al., 2012; Mírovskỳ et al., 2012), this approach takes advantage of various unique features of the t-layer, the most important of which is probably the reconstruction of elided verbs in constructions such as *Marie odjela a Lucie také*. [*Mary has gone and so has Lucy,* lit. *Mary has gone and Lucy too.*].

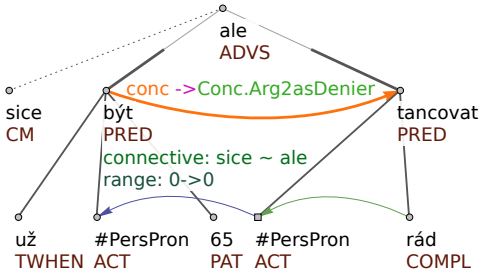

Figure 1: A discourse relation on top of a tectogrammatical tree, represented by the thick orange arrow.

Figure 1 shows a discourse relation from Example 1 depicted by the thick orange arrow in a tectogrammatical tree representation of the sentence. The arrow connects roots (*representative nodes*) of the subtrees that correspond to the two arguments. It goes from the *start node* of the relation to the *target node*. The direction of the arrow determines the argument semantics: in the case of *concession*, the arrow goes from the argument that raises an expectation to the argument that brings the denial.

The annotation is captured in the Prague Markup Language data format (PML; Hana and Štěpánek, 2012), an XML-based format for complex multi-layer linguistic annotations. PML is a general XML-based format accompanied by an application framework for complex multi-layer linguistic annotations, with tools available for browsing and editing the data (tree editor TrEd;[8] Pajas and Štěpánek, 2008, conf. Figure 1), for script-processing the data (a command-line tool `btred`), and for graphically oriented and powerful searching in the data (PML-TQ;[9] Pajas and Štěpánek, 2009).

All information about a single discourse relation is kept at its start node in several sub-attributes of a structured attribute *discourse*. Its sub-attribute *target_node.rf* contains an *id* (a corpus-wide unique identifier) of the target node. Sub-attributes *discourse_type* and *sense_PDTB3* carry the discourse type and the sense of the relation. Sub-attributes *t-connectors.rf* and (if needed) *a-connectors.rf* carry lists of nodes (on two annotation layers) that represent the connective of the relation. For relations marked by secondary connectives, sub-attribute *is_secondary* is set to *1*. Sub-attributes *start_range* and *target_range* carry information about the extent of the two arguments.

Typically, the extent of an argument of a discourse relation in PDiT corresponds to the subtree of the representative (start or target) node, with the

[8] `https://ufal.mff.cuni.cz/tred/`

[9] Prague Markup Language – Tree Query, `https://ufal.mff.cuni.cz/pmltq`

exclusion of the other argument in subordinated constructions. These typical cases are indicated in the respective range attribute by value *0*. Less frequently, the argument also includes one or more subsequent sentences (the number of sentences is then given in the respective range attribute). In rare cases, complex arguments need to be specified by arbitrarily definable groups of nodes – in that case, each node carries information about its membership in one or more groups, the range sub-attribute contains the value *group* and the identifier of the group is carried in the sub-attribute *start_group_id*, or *target_group_id*, respectively.

The PDiT annotation is available in two data formats: (i) in the original PML data format, as shortly described above, (ii) in an automatic transformation to the PDTB 3.0 stand-off column text format. In the column format, a discourse relation is represented by a single line consisting of a number of fields separated with '|', with each field carrying a single piece of annotation information; several fields provide links (offsets) to plain text versions of the documents. Fields 0–33 correspond to the original PDTB 3.0 fields and ensure the compatibility with PDTB-related tools; fields 34–43 are unique to PDiT and carry additional information, such as the original discourse type, textual representation of several previous fields, and (for the PDT subcorpus) the genre of the document (see Mírovský et al., 2023 for details on the transformation process).

2.2. CzeDLex

From PDiT 3.0 on, Czech Lexicon of Discourse Connectives (CzeDLex, Mírovský et al., 2017)[10] has become an integral part of discourse relations annotation both in automatic and manual procedures. It is used both as a part of a rule-based discourse parser and as a consistency-supporting tool during manual annotations. Based mainly on PDiT 2.0 and subsequently enriched with data from other sources, CzeDLex includes approx. 2 hundred level-one entries (e.g. *když* [*when*], *tak* [*so*], most of them also covering numerous complex forms (e.g. *když ~ tak* [*when ~ so*]) or modifications (*právě když* [*exactly when*]). All entries are provided with complex linguistic information (part-of-speech characteristics, argument semantics, word order) as well as corpus characteristics (number of total usages, usages in various discourse types, corpus examples).

3. PDiT 4.0 Annotation Process

The Prague Discourse Treebank 4.0 brings discourse annotation to all four subcorpora of the underlying corpus, the Prague Dependency Treebank - Consolidated 2.0, which are:

1. the original Prague Dependency Treebank (PDT),
2. the Czech part of the Prague Czech-English Dependency Treebank (PCEDT-cz),
3. the Prague Dependency Treebank of Spoken Czech (PDTSC), and
4. the Faust corpus

In four parts of this section, we describe discourse relations annotation process for each of the four subcorpora, which differ not only in size and text types but also in the annotation method and complexity. The PDT subcorpus was annotated practically entirely manually, building mostly upon the previous versions (Poláková et al., 2012; Rysová et al., 2016; Synková et al., 2022). For the PCEDT-cz part, annotation projection from PDTB 3.0 was combined with automatic discourse parsing, followed by manual discrepancy checks (Mírovský et al., 2024). The PDTSC and Faust subcorpora were annotated using automatic discourse parsing with extensive manual verification.

Table 2 gives an overview of the total size of PDiT 4.0 and its four subcorpora, and the number of annotated discourse relations. Table 3 summarizes the annotation method and quality for the individual subcorpora.

3.1. Prague Dependency Treebank (PDT)

3.1.1. Data

The PDT part of PDiT 4.0 covers the data of the Prague Discourse Treebank 3.0, i.e. the Prague Dependency Treebank annotated fully manually with discourse relations. The Prague Dependency Treebank consists of written journalistic texts originating from the nineties of the 20th century. It has slightly more than 830 thousand tokens in approx. 49 thousand sentences.

3.1.2. Annotation process

The discourse annotation in PDT started back in 2009, culminating in 2012 with the publication of the first version of PDiT. From the very beginning, the process was based on manual annotation. Annotators read the whole text, marking discourse connectives first on paper, and then they annotated the discourse relations in tree editor TrEd directly on top of the deep syntax (tectogrammatical) trees. They were aware of linguistic information already captured at the tectogrammatical layer (t-layer)[11] and thus relations that could be

[10] https://ufal.mff.cuni.cz/czedlex/

[11] which was also annotated manually

	corpus size (documents)	corpus size (sentences[a])	corpus size (tokens)	discourse relations (PML format)	discourse relations (PDTB format[c])
PDT	3,165	49,419	833,195	21,611 (+ 443 list[b] rels)	21,537
PCEDT-cz	2,312	49,208	1,152,289	28,967	28,940
PDTSC	1,553	73,802	742,316	31,218	31,074
Faust	60	3,000	33,836	710	705
TOTAL	7,090	175,429	2,761,636	82,506 (+ 443 list[b] rels)	82,256

Table 2: Subcorpora sizes and numbers of discourse relations in individual PDiT 4.0 subcorpora

[a] In the PML version of the PDT and PDTSC data, there are 9 and 33 empty trees, respectively. These empty trees are not included in the numbers of sentences in the corpora. (Including the empty trees, the numbers of sentences in the PDT and the PDTSC would be 49,428 and 73,835, respectively. The total number of sentences in PDT-C (i.e, in the PML version of PDiT) would then be 175,471).

[b] List rels are relations connecting items of lists (e.g. enumerated items 1), 2)) and as such they represent another type of discourse phenomena than semantic discourse relations.

[c] The tectogrammatical annotation reconstructs several types of ellipses, which sometimes creates a duplication of an intra-sentential discourse relation; mainly for this reason and also for some theory incompatibilities between the Prague and Penn approaches, the number of discourse relations in the PDTB format is smaller than in the PML format.

obtained from this layer without any loss of information (e.g., conditional clauses with semantic label COND that express discourse relation *condition* (as opposed to those expressing *pragmatic condition*)) were left for automatic post-processing (for details, see Jínová et al., 2012).

In subsequent versions, this annotation was gradually improved: errors were fixed, secondary connectives were included (see details in Rysová and Rysová, 2015), genres of documents were annotated (Poláková et al., 2014). In the version 3.0, the PDiT discourse annotation was made consistent with the lexicon of connectives CzeDLex, annotation of pragmatic relations was revised and the data were transformed into the PDTB format and taxonomy. For PDiT 4.0, the annotation was adjusted to the updated underlying tectogrammatical annotation and a new projection to the PDTB format and taxonomy was performed.

3.1.3. Annotation quality

The inter-annotator agreement in the PDT part of PDiT was measured on 2 thousand double-annotated sentences (4% of the data) during the original annotation of PDiT 1.0, using the connective-based F1-measure (Mírovský et al., 2010) for measuring the agreement on the recognition of a discourse relation. A simple ratio and Cohen's κ were used for measuring the agreement on the discourse type of relations recognized by both annotators. According to Poláková et al. (2013), the agreement on relation recognition achieved the F1 score of 0.83, while discourse type classification reached 77% accuracy (Cohen's kappa 0.71).[12]

These numbers do not reflect extensive annotation quality and consistency improvements conducted in subsequent years and published in numerous subsequent minor and major updates of PDiT. However, no large-scale measurements of the quality of the updated annotations have been done.[13] Despite the lack of recent numerical quality checks, we consider the PDT part of PDiT to be of the highest reliability level among all four PDiT parts, thanks to the completely manual annotation based on reading all the texts, and thanks to the number of corrections in subsequent years.

The discourse annotation of the PDT stands apart from the annotation of the three other PDiT parts also in its broadness. Only the PDT part contains a systematic annotation of secondary connectives (which were marked only marginally in the other parts). Also genres of documents, headings, metatextual notes and captions, as well as list relations (i.e. relations structuring text by enumerations, see 443 list relations in Table 2), are only annotated in the PDT part of PDiT 4.0.

[12] For comparison, the simple ratio agreement on types in discourse relations (77%) is the closest measure to the way of measuring the inter-annotator agreement used on subsenses in the Penn Discourse Treebank 2.0, reported in Prasad et al. (2008). Their agreement was 80%.

[13] among other reasons because the checks and updates were done by only one expert annotator

3.2. Prague Czech-English Dependency Treebank, Czech part (PCEDT-cz)

3.2.1. Data

The Prague Czech-English Dependency Treebank 2.0 (PCEDT; Hajič et al., 2012) is a manually annotated parallel treebank sized over 1.1 million tokens in approx. 49 thousand sentences for each language part. The English side of the PCEDT 2.0 covers the Wall Street Journal section of the Penn Treebank (Marcus et al., 1995), which are mainly texts with economic themes, including reports on the development of the stock market.

The Czech side of the data, the PCEDT-cz, is a subject for discourse annotation within PDiT 4.0. It consists of manual Czech translations of all of the Penn Treebank-WSJ texts with (by design) 1:1 sentence alignment, annotated up to the tectogrammatical layer.

3.2.2. Annotation process

The PCEDT-cz discourse annotation process is described in detail in Mírovský et al. (2024). It was based on combination of annotation projection from the PDTB 3.0[14] and automatic Czech discourse parsing.[15] Discourse relations resulting from the two methods (henceforth called "projected" and "automatic") were subsequently partially automatically processed and, depending on their overlap, merged.

The automatic merge could be done if the projected and automatic arrows overlapped perfectly (they shared the same start and target nodes), had the equivalent type and sense (as shown by Table 1) and had word-aligned connectives. Contexts in which any of these conditions were violated were analyzed in detail on data samples and sets of rules were devised to also process parts of these cases automatically.[16]

Even after applying all manually devised rules, many relations remained in the data that could not be automatically merged without compromising the annotation quality. These relations (both projected and automatic) were then checked by experienced annotators and processed manually. The manual interventions comprised of: decisions on existence of relations (in approx. 6 thousand positions), discourse types and senses (in approx. 8 thousand positions), annotations of new relations (approx. 850 relations). As a final step, all discourse relations were transformed to corresponding PDTB senses.

3.2.3. Annotation quality

A thousand sentences were annotated completely manually, as a basis for measuring the annotation quality. The first measurement was performed purely on the output of the discourse parser, giving these results: F1 on relation presence: 0.85 (0.64 for inter-sentential, 0.35 on exact argument extent), agreement on discourse types: 74% (Cohen's kappa 0.68).[17] The second measurement, to evaluate the contribution of the annotation projection and manual interventions, was performed on the final data: F1 on relation presence: 0.87 (0.75 for inter-sentential, 0.46 on exact argument extent), agreement on discourse types: 78% (Cohen's kappa 0.73).

3.3. Prague Dependency Treebank of Spoken Czech (PDTSC)

3.3.1. Data

The Prague Dependency Treebank of Spoken Czech (PDTSC; Mikulová et al., 2017) is a corpus of texts representing over 120 hours of spontaneous dialogues. It consists of more than 742 thousand tokens in almost 74 thousand sentences, coming from two sources: (i) slightly moderated dialogs with Holocaust survivors, and (ii) dialogs between an avatar on the computer screen[18] and someone reminiscing about personal photograph collections. The texts have been manually transcribed from the audio layer, reconstructed to comply with Czech syntax,[19] and subsequently annotated up to the tectogrammatical layer.

3.3.2. Annotation process

In contrast to the PCEDT-cz part, the annotation process in the PDTSC part consisted only of two steps. First, the whole PDTSC data were annotated fully automatically by the Czech discourse parser. The rule-based discourse parser utilizes information from the (manually annotated) tectogrammatical layer, partially also from the surface

[14] taking advantage of the fact that the PDTB annotates the same texts as the English part of the PCEDT

[15] Some details on the parser are given in 3.3.2.

[16] For example, intra-sentential projected and automatic relations sharing both start and target nodes, but disagreeing in one of the other aspects, were merged automatically when there were no other relations sharing either the start or target nodes with them.

[17] The good results of the parser come largely from the fact that the underlying tectogrammatical layer was annotated manually.

[18] acting as an autonomous system but – without the participants' knowledge – actually controlled by a human (Wizard-of-Oz setup)

[19] The reconstruction process removed repetitions, corrections and other speech events that would disrupt the syntactic correctness of the sentence, see Hajič et al. (2008).

	annotation method (initially, checks) intra-sentential	inter-sentential	annotation quality (F1, accuracy, Cohen's κ) method vs. human	human vs. human
PDT	tectogrammatics	manual	-	0.83, 77%, 0.71[a]
checks	all			
PCEDT-cz	parser + projection	parser + projection	0.87, 78%, 0.73	-
checks	all relevant	all relevant		
PDTSC	parser	parser	0.94, 83%, 0.8	-
checks	all relevant	all		
Faust	parser	parser (rare)	-	-
checks	all	all		

Table 3: Annotation method and quality measurement in individual PDiT 4.0 subcorpora. In the two annotation method columns, the first row for each subcorpus describes the first annotation step, the second row describes the extent of subsequent manual checks.

[a] only measured in the first version of PDiT in 2013, not reflecting subsequent annotation and checks

syntax layer (at that time not manually annotated), and from the lexicon of Czech discourse connectives CzeDLex. The tectogrammatical tree structure and the deep syntax labels (functors), along with information from CzeDLex, are employed to determine the existence of a relation, locate the connective, classify the discourse type[20] and determine the arguments extent.

As PDTSC represents spoken language, at the beginning of the annotations, the parser was adjusted for its specific phenomena. In an iterative process, the results of the automatic annotation on data samples were analyzed manually and the parser was gradually improved according to the detected problems (e.g., the parser was allowed to search for connective nodes deeper in the tectogrammatical tree, the lexicon CzeDLex was enriched with new entries).

The remaining part of the annotation encompassed numerous types of manual revisions of the automatic annotation.

Types of manual checks: The manual analysis of samples of the parser output revealed the most frequent types of errors that could not be (with a reasonable effort) solved by further improving the parser. Naturally, more discrepancies were found in annotation of inter-sentential relations, as the annotation of the tectogrammatical layer helps solve most of ambiguities in the intra-sentential relations. Delimitation of the discourse argument extent is one of such issues. Manual checks led to corrections of the argument extent in 3.5 thousand inter-sentential relations. For intra-sentential relations, much fewer positions needed to be checked with about 600 performed changes.

Manual revisions proceeded from uncommon combinations of connectives and discourse types (e.g., connective *místo aby* [*instead of*] with discourse type *conjunction* (instead of the expected *correction*[21])) and highly ambiguous connectives (e.g., *když* [*when*, *as*, *if*, *while*] that can signal *condition*, *precedence-succession*, *synchrony*, *reason-result*, *specification* and other discourse types) to more common combinations and less (but still) ambiguous connectives. Again, without the clues on the tectogrammatical annotation layer, the inter-sentential cases required more effort.

Connectives not existing in CzeDLex (but still somehow detected by the parser) were all checked (approx. 15 cases, e.g., *z* [here: *for*] instead of *z důvodu* [*for the reason*]), as well as all 27 cases in which the parser failed to identify any discourse type.

Further, expressions homonymous with connectives were often erroneously marked by the parser to have the connective function, so approx. 6 thousand such relations were deleted.

Discourse types requiring special attention: Special attention was paid to the *opposition* relation, as the parser could not distinguish all subtypes of contrastive relations (*restrictive opposi-*

[20] The functors help determine the discourse type for intra-sentential relations. For example, the functor COND is used for conditional meaning and when a node with COND functor is occupied by a verb, it is clear that it represents a root of a conditional dependent clause and the parser can mark *condition* between this node and a node of the governing clause. There are also functors indicating reason or time meaning of dependent clauses and other functors that help identify discourse relations in coordinated structures. For inter-sentential relations, the parser relies mostly on the lexicon CzeDLex.

[21] *Correction* is a relation between a denied statement and another one that replaces it: *…stráže SS* ne*byly na svých místech.* **Byly tam** ale **jiné uniformy.** [*...the SS guards were* not *at their posts.* But **there were other uniforms there.**]

tion,[22] *correction*, *concession*, *confrontation*, *opposition*) when they were signalled by non-specific connectives (e.g., *ale* [*but*] can signal all types of the contrastive relations).

In the first step, contexts with possible formal signals for *correction* (such as negation in the first part of a coordinated structure with the *ADVS* functor) and *restrictive opposition* (particles *jen*, *jenom*, *pouze* [*just*, *only*] in the second part of a coordinated structure with the same functor) were checked manually. In the second step, the rest of both inter- and intra-sentential *opposions* were also checked manually. In total, almost 1.4 thousand *opposition* relations (out of 4.2 thousand) were converted to a more specific discourse type.

The most frequently re-annotated type, however, was *precedence-succession* (almost 1.6 thousand re-annoted cases out of 5.1 thousand automatically detected ones). On the one hand, the tectogrammatical layer does not help distinguish this relation from *synchrony* and, on the other hand, connectives typical for *precedence–succession* often signal also *condition* or *reason–result* and there is no formal clue that would help distinguish these cases automatically.

Special attention needed to be paid also to pragmatic relations, as they are more frequent in originally spoken texts than in written ones. Pragmatic relations were defined as (i) relations that hold between presuppositions or another pragmatic phenomenon is involved, (ii) relations where the form and the meaning do not correspond (but at the same time the relation cannot be interpreted as another semantic relation), including stylistically inappropriate contexts (see more in Synková et al., 2024).

The only formal clue we could find for distinguishing pragmatic relations from "normal" ones was for *condition* (combination of parenthesis and a conditional connective) and such cases covered only a small portion of *pragmatic conditions* in the texts. Therefore, all *opposition*, *reason–result* and *condition* relations were manually checked to detect their pragmatic counterparts. In sum, these checks, as well as revisions of non-typical combinations of a connective and a discourse type, revealed 400 *pragmatic reason–results*[23], 165 *pragmatic oppositions*[24] and 150 *pragmatic conditions*[25] (for comparison, PDiT 3.0 contains approx. 60 *pragmatic reason–results*, 30 *pragmatic oppositions* and 100 *pragmatic conditions*).

It is worth noting that manual checks also led to a significant increase of relations *generalization* (350 cases in the revised data as opposed to 150 cases in the parser output) and *equivalence* (320 vs. 40). This increase reflects the fact that (i) the connectives of these relations are homonymous with connectives of much more frequent relations (and the parser preferred the more frequent types) and (ii) these relations are more frequent in spoken than written data.

Connective changes: Relations at a given place were always checked comprehensively – in addition to the presence of a relation, the extent of arguments and the discourse type, also the connective was checked. Corrections of connectives were relatively less numerous compared to other types of changes (1.4 thousand in total), the most frequent being (i) addition of negation (mostly to *ale* [*but*]), (ii) addition of the second part of a correlative connective, and (iii) addition of a modification particle (such as *hlavně* [*mainly*], *jen* [*only*], etc.).

Annotation of new relations: Apart from extensive checks of existing relations, revisions of selected contexts led to annotation of new relations not included in the automatic annotation. These relations often occurred in contexts with several connectives for more than one relation between the same arguments. The most frequent types for newly annotated relations were *precedence–succession* (368 cases), followed by *conjunction* (230), *opposition* (134) and *correction* (123). The *precedence–succession* connectives are the ones most often combined with connectives of other relations (typically *ale pak* [*but then*]). Moreover, connectives for *precedence–succession* often occur in non-typical positions within the sentence and the parser sometimes failed to identify them as connectives.

[22] *Restrictive opposition* is a relation between a statement and an exception or constraint to it, as illustrated by sentences "Jak už jsem předtím říkal, *vypomáhala finančně tím, že doma občas známým a příbuzným něco ušila*. **Byla to** ale **jenom taková výpomoc v uvozovkách**. [As I said before, *she helped out financially by occasionally sewing things for friends and relatives at home*. But **it was only a little help, so to speak**.]"

[23] *Pragmatic reason–result* occurs e.g. in the sentence *Asi jsem trošku zlobila,* protože **dokonce ještě na ekonomické škole jsem měla dvojku z chování...** [*I must have been a bit naughty,* because **even at business school I got a B for behavior.**] where a subjective belief is expressed.

[24] *Pragmatic opposition* can be illustrated e.g. by the sentence *Nebudete tomu věřit*, ale **jezdím se svým synem a s jeho rodinou již šestý rok společně na dovolenou.** [*You won't believe it,* but **I've been going on vacation with my son and his family for six years now.**] where there is no contrast between the content of clauses, but the first clause sets the stage for the acceptance of the second clause.

[25] *Pragmatic condition* often has a form of a conditional dependent clause that semantically comments on the certainty with which the content of the main clause applies, as illustrated by this example: "*To znamená, že jsem tam byl v roce 1995,* jestli **dobře počítám.** [*That means I was there in 1995,* if **I'm counting correctly.**]

Most new relations were annotated as a result of reading contexts where automatic relations occurred, but places where relations could be omitted by the parser were also searched systematically: (i) special constructions (e.g., the second part of coordinated structures or in relative clauses), (ii) expressions homonymous with most frequent connectives not added to discourse relations as connectives were checked to find out whether some discourse relations with more complicated nature were present in the data. This way, however, only a few new relations were detected.

Annotation revisions summary: In total, the PDTSC contained more than 36 thousand automatically annotated relations. More than 11 thousand inter-sentential relations were checked manually, as well as more than 17 thousand intra-sentential relations. Whereas 6 thousand of these relations were deleted, 1.2 thousand new relations were added completely manually. In the final version, the PDTSC contains 31 thousand discourse relations. There are 7.4 thousand inter-sentential relations (which were all manually checked) and 23.6 thousand intra-sentential relations (out of which 8.4 thousand were reliable enough not to require manual checks). The discourse type was revised in 6.4 thousand cases, the argument extent was changed in 4.1 thousand cases, and the connective was modified in 1.4 thousand cases. As a final step, all discourse relations were transformed to corresponding PDTB senses.

3.3.3. Annotation quality

The annotation quality was again measured twice, on a thousand completely manually annotated sentences. The parser itself performed with F1 on relation presence 0.89 (0.7 for inter-sentential, 0.29 on exact argument extent), with agreement on discourse types 76% (Cohen's kappa 0.71). The measurement on the final data gave F1 on relation presence 0.94 (0.85 for inter-sentential, 0.45 on exact argument extent) and agreement on discourse types 83% (Cohen's kappa 0.8).

3.4. PDT-Faust

PDT-Faust is a small part of the PDT-C, compiled out of short, often non-standard pieces of text (mostly one or only a few sentences long), collected as a user-generated input during the development of an automatic web translator. It consists of 34 thousand tokens in 3 thousand sentences. As the short text segments do not form a coherent text, most of the discourse relations here are intra-sentential.[26]

The annotation process consisted of the same steps as for the PDTSC, i.e. (i) automatic discourse parsing, (ii) manual checks, (iii) transformation into PDTB senses.

Given the small size, all automatically parsed relations were checked manually. Subsequently, all expressions with a possible discourse signalling function but omitted by the discourse parser were manually checked and annotated where appropriate. PDT-Faust contained 720 automatically annotated discourse relations; after the manual checks, the released version contains 710 relations. Given the annotation procedure, the PDT-Faust data can be considered fully manually discourse-annotated.

4. Conclusion

The Prague Discourse Treebank 4.0 (PDiT 4.0) is a large genre-diversified corpus of Czech texts annotated with explicit discourse relations. In more than 2.7 million tokens (175 thousand sentences), 82.5 thousand discourse relations have been annotated. Some parts of the corpus were annotated completely manually, other parts using cost-effective partially automatic methods, with comparable annotation quality.

Together with the Penn Discourse Treebank 3.0, the PCEDT-cz part of PDiT 4.0 (1.2 million tokens in 49 thousand sentences) represents parallel discourse annotation of parallel texts in the Prague Czech–English Dependency Treebank.

The Prague Discourse Treebank 4.0[27] was published under the Creative Commons licence in December 2024 and is available from the LINDAT-CLARIAH-CZ repository in two data formats:

- PDiT 4.0 in its native PML data format, as a part of the PDT-C 2.0 (Hajič et al., 2024), i.e. all underlying annotation layers + discourse annotation as a part of the tectogrammatical layer,[28]
- PDiT 4.0 in the PDTB 3.0 data format (Synková et al., 2024), i.e. raw texts + stand-off discourse annotation.[29]

All discourse relations (in both data formats) were converted to align with the sense taxonomy of the Penn Discourse Treebank 3.0, ensuring compatibility with this widely recognized framework and facilitating accessibility for researchers.

[26] The discourse annotation of the PDT-Faust part of the PDT-C has been done mainly in the interest to complete the discourse annotation in the whole PDT-C. Given the incoherence of the texts (and subsequent lack of inter-sentential relations), the discourse annotation of this small corpus is of limited importance.

[27] https://ufal.mff.cuni.cz/pdit4.0

[28] http://hdl.handle.net/11234/1-5813

[29] http://hdl.handle.net/11234/1-5680

Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project 22-03269S) and the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2023062).

Appendix A: Fields description of the adapted PDTB 3.0 data format

Table 4 below describes the PDTB 3.0-compatible column data format used in PDiT 4.0. Each record consists of 44 fields. Fields 0–33 correspond to fields defined in the PDTB 3.0 and their description is taken from the PDTB 3.0 annotation manual[30] (and slightly adapted for PDiT 4.0), fields 34–43 carry additional information added only in PDiT 4.0. Some of the fields are not used in PDiT 4.0 but they are kept for compatibility and completeness (they are marked with grey background).

Appendix B: Frequencies of discourse types and senses in PDiT 4.0

Table 5 below shows frequency of various discourse types annotated in individual parts and in total in PDiT 4.0, measured in the PML data format and sorted by the frequency in the whole corpus.

Table 6 shows frequency of various PDTB 3.0 senses annotated in individual parts and in total in PDiT 4.0, measured in the PDTB 3.0 data format and sorted by the frequency in the whole corpus.

5. Bibliographical References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, et al. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2727–2734. European Language Resources Association (ELRA).

Amal Al-Saif and Katja Markert. 2010. The leeds arabic discourse treebank: Annotating discourse connectives for arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Peter Bourgonje and Manfred Stede. 2020. The potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1061–1066.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. 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.

Jan Hajič, Silvie Cinková, Marie Mikulová, Petr Pajas, Jan Ptáček, Josef Toman, and Zdeňka Urešová. 2008. PDTSL: An annotated resource for speech reconstruction. In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, pages 93–96, Goa, India. IEEE.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Eva Hajičová, Jarmila Panevová, and Petr Sgall. 2008. Tectogrammatics in corpus tagging. In *Perspectives on Semantics, Pragmatics, and Discourse*, pages 293–299. John Benjamins Publishing Company.

Jirka Hana and Jan Štěpánek. 2012. Prague markup language framework. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 12–21, Stroudsburg, PA, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Pavlína Jínová, Jiří Mírovskỳ, and Lucie Poláková. 2012. Semi-automatic annotation of intrasentential discourse relations in pdt. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects*, pages 43–58.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Gabriel Murray. 2019. Discourse analysis and its applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 12–17, Florence, Italy. Association for Computational Linguistics.

[30] https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf

Index	Field Name	Description
0	Relation Type	Explicit, AltLex, AltLexC
1	Conn SpanList	SpanList of the Explicit Connective or the AltLex/AltLexC selection
2	Conn Src	Connective's Source
3	Conn Type	Connective's Type
4	Conn Pol	Connective's Polarity
5	Conn Det	Connective's Determinacy
6	Conn Feat SpanList	Connective's Feature SpanList
7	Conn1	Explicit Connective Head
8	SClass1A	Semantic Class of the Connective
9	SClass1B	Second Semantic Class of the First Connective
10	Conn2	Second Implicit Connective
11	SClass2A	First Semantic Class of the Second Connective
12	SClass2B	Second Semantic Class of the Second Connective
13	Sup1	SpanList SpanList of the First Argument's Supplement
14	Arg1	SpanList SpanList of the First Argument
15	Arg1 Src	First Argument's Source
16	Arg1 Type	First Argument's Type
17	Arg1 Pol	First Argument's Polarity
18	Arg1 Det	First Argument's Determinacy
19	Arg1 Feat SpanList	SpanList of the First Argument's Feature
20	Arg2 SpanList	SpanList of the Second Argument
21	Arg2 Src	Second Argument's Source
22	Arg2 Type	Second Argument's Type
23	Arg2 Pol	Second Argument's Polarity
24	Arg2 Det	Second Argument's Determinacy
25	Arg2 Feat SpanList	SpanList of the Second Argument's Feature
26	Sup2 SpanList	SpanList of the Second Argument's Supplement
27	Adju Reason	The Adjudication Reason
28	Adju Disagr	The type of the Adjudication disagreement
29	PB Role	The PropBank role of the PropBank verb
30	PB Verb	The PropBank verb of the main clause of this relation
31	Offset	The Conn SpanList of Explicit/AltLex/AltLexC tokens
32	Provenance	Indicates whether the token is a new PDTB3 token or has a corresponding PDTB2 token
33	Link	The link id of the token
34	Discourse Type	The original discourse type in the Prague taxonomy
35	Conn Text	Text representation of field 31 (Offset)
36	Conn Feat Text	Text representation of field 6 (Conn Feat SpanList)
37	Sup1 Text	Text representation of field 13 (Sup1 SpanList)
38	Arg1 Text	Text representation of field 14 (Arg1 SpanList)
39	Arg1 Feat Text	Text representation of field 19 (Arg1 Feat SpanList)
40	Arg2 Text	Text representation of field 20 (Arg2 SpanList)
41	Arg2 Feat Text	Text representation of field 25 (Arg2 Feat SpanList)
42	Sup2 Text	Text representation of field 26 (Sup2 SpanList)
43	Genre	The genre of the document

Table 4: Field definitions in the Prague Discourse Treebank 4.0 corresponding to the fields defined in the PDTB 3.0 (fields 0–33) and additional fields (34–43) present in the PDiT 4.0 column data format. Fields not used in PDiT 4.0 are highlighted with grey background.

discourse type	PDT	PCEDT-cz	PDTSC	Faust	Total
conjunction	7,723	11,140	10,388	258	29,509
reason–result	3,031	2,712	5,087	97	10,927
opposition	3,192	4,103	3,017	36	10,348
precedence–succession	1,031	1,845	3,652	72	6,600
condition	1,331	1,921	1,460	89	4,801
synchrony	260	889	1,638	12	2,799
concession	897	1,172	645	22	2,736
purpose	421	1,524	726	39	2,710
confrontation	690	1,055	675	12	2,432
specification	677	539	209	3	1,428
correction	423	431	520	16	1,390
gradation	463	268	375	2	1,108
restrictive opposition	289	162	559	9	1,019
conjunctive alternative	100	357	385	12	854
instantiation	208	351	128		687
disjunctive alternative	267	80	266	10	623
equivalence	127	90	323	4	544
generalization	137	32	353		522
pragmatic reason–result	61	26	400	6	493
explication	147	128	93	5	373
pragmatic condition	106	106	154	6	372
pragmatic opposition	30	36	165		231
(list relation)	(443)				(443)
Total (without list relations)	21,611	28,967	31,218	710	82,506

Table 5: Frequency of discourse types in individual parts and in total in the Prague Discourse Treebank 4.0, measured in the PML data format and sorted by the frequency in the whole corpus. (Please see footnote *c* at Table 2 for explanation of the difference between the counts of relations in the PML and PDTB 3.0 data formats.)

Ai Kubota, Takuma Sato, Takayuki Amamoto, Ryota Akiyoshi, and Koji Mineshima. 2024. Annotation of japanese discourse relations focusing on concessive inferences. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1215–1224.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8:243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1995. *Treebank-2*. Data/Software, Linguistic Data Consortium. University of Pennsylvania, Philadelphia. LDC95T7.

Amália Mendes and Pierre Lejeune. 2022. Crpc-db a discourse bank for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 79–89. Springer.

Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jan Štěpánek, and Jan Hajič. 2017. PDTSC 2.0 - spoken corpus with rich multi-layer structural annotation. In *20th International Conference, TSD 2017 Prague, Czech Republic, August 27–31, 2017 Proceedings*, number 10415 in Lecture Notes in Computer Science, pages 129–137, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Jiří Mírovskỳ, Pavlína Jínová, and Lucie Poláková. 2012. Does tectogrammatics help the annotation of discourse? In *Proceedings of COLING 2012: Posters*, pages 853–862.

Jiří Mírovský, Lucie Mladová, and Šárka Zikánová. 2010. Connective-based measuring of the inter-annotator agreement in the annotation of discourse in PDT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, volume 1, pages 775–781, Beijing, China. Chinese Information Processing Society of China, Tsinghua University Press.

Jiří Mírovský, Magdaléna Rysová, Pavlína Synková, and Lucie Poláková. 2023. Prague

sense	PDT	PCEDT-cz	PDTSC	Faust	Total
Expansion.Conjunction	8,133	11,330	10,753	260	30,476
Comparison.Concession.Arg2-as-denier	3,539	4,448	3,479	45	11,511
Contingency.Cause.Reason	1,752	1,870	2,897	64	6,583
Contingency.Cause.Result	1,306	850	2,381	33	4,570
Contingency.Condition.Arg2-as-cond	1,238	1,771	1,407	76	4,492
Temporal.Asynchronous.Precedence	686	813	2,825	33	4,357
Temporal.Synchronous	260	889	1,627	12	2,788
Contingency.Purpose.Arg2-as-goal	415	1,518	723	39	2,695
Comparison.Contrast	784	1,082	697	12	2,575
Temporal.Asynchronous.Succession	345	1,032	810	39	2,226
Comparison.Concession.Arg1-as-denier	564	916	435	13	1,928
Expansion.Disjunction	366	435	651	21	1,473
Expansion.Level-of-detail.Arg2-as-detail	650	558	258	3	1,469
Expansion.Substitution.Arg2-as-subst	368	225	459	13	1,065
Expansion.Instantiation.Arg2-as-instance	206	351	129		686
Expansion.Equivalence	127	90	324	4	545
Expansion.Level-of-detail.Arg1-as-detail	136	33	359		528
Expansion.Exception.Arg2-as-excpt	196	44	185	9	434
Contingency.Condition+SA	102	101	136	6	345
Expansion.Substitution.Arg1-as-subst	55	208	61	3	327
Contingency.Cause+Belief.Reason+Belief	123	104	63	6	296
Expansion.Exception.Arg1-as-excpt	9	13	137		159
Contingency.Negative-condition.Arg2-as-negCond	47	90	7	3	147
Contingency.Condition.Arg1-as-cond	47	53	26	5	131
Comparison.Similarity	47	73	1		121
Comparison.Concession+SA.Arg2-as-denier+SA	5	21	77		103
Contingency.Cause+SA.Result+SA	4		82		86
Contingency.Cause+SA.Reason+SA	3	4	54		61
Contingency.Cause+Belief.Result+Belief	7	7	13	5	32
Contingency.Negative-condition.Arg1-as-negCond	2	5	16	1	24
Contingency.Purpose.Arg1-as-goal	6	5	1		12
Contingency.Negative-cause.NegResult	7				7
Expansion.Instantiation.Arg1-as-instance	2		1		3
Contingency.Negative-condition+SA		1			1
Total	21,537	28,940	31,074	705	82,256

Table 6: Frequency of PDTB 3.0 senses in individual parts and in total in the Prague Discourse Treebank 4.0, measured in the PDTB 3.0 data format and sorted by the frequency in the whole corpus. All *SpeechAct* labels have been shortened to *SA*. (Please see footnote *c* at Table 2 for explanation of the difference between the counts of relations in the PML and PDTB 3.0 data formats.)

to penn discourse transformation. *The Prague Bulletin of Mathematical Linguistics*, (120):5–30.

Jiří Mírovský, Pavlína Synková, Lucie Polakova, and Marie Paclíková. 2024. Cost-effective discourse annotation in the Prague Czech–English Dependency Treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4067–4077, Torino, Italia. ELRA and ICCL.

Jiří Mírovský, Pavlína Synková, Lucie Poláková, and Marie Paclíková. 2024. Cost-effective discourse annotation in the prague czech–english dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4067–4077, Torino, Italy. European Language Resources Association.

Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, and Lucie Poláková. 2017. CzeDLex – a lexicon of czech discourse connectives. *The Prague Bulletin of Mathematical Linguistics*, (109):61–91.

Verginica Barbu Mititelu and Tudor Voicu. 2024. Function multiword expressions annotated with

discourse relations in the romanian reference treebank. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 90–97.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Suman Meena, Dipti Misra Sharma, and Aravind Joshi. 2009. Experiments with annotating discourse relations in the hindi discourse relation bank. In *Proceedings of the 7th International Conference on Natural Language Processing (ICON-2009), Hyderabad, India*.

Petr Pajas and Jan Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester. The Coling 2008 Organizing Committee.

Petr Pajas and Jan Štěpánek. 2009. System for querying syntactically annotated corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, Singapore. Association for Computational Linguistics.

Lucie Poláková, Pavlína Jínová, and Jirí Mírovskỳ. 2014. Genres in the prague discourse treebank. In *LREC*, pages 1320–1326.

Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzana Bedřichová, Jiří Mírovskỳ, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, and Eva Hajičová. 2012. Manual for annotation of discourse relations in the prague dependency treebank. *Technical report. Charles University, Prague*, 26(6).

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. 2013. Introducing the prague discourse treebank 1.0. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan. Asian Federation of Natural Language Processing, Asian Federation of Natural Language Processing.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2961–2968.

Ponrawee Prasertsom, Apiwat Jaroonpol, and Attapol T Rutherford. 2024. The thai discourse treebank: Annotating and classifying thai discourse connectives. *Transactions of the Association for Computational Linguistics*, 12:613–629.

Magdaléna Rysová and Kateřina Rysová. 2015. Secondary connectives in the prague dependency treebank. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 291–299.

Pavlína Synková, Jiří Mírovskỳ, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the prague discourse treebank 3.0. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1270–1279.

Deniz Zeyrek, Isin Demirsahin, Ayisigi B Sevdik-Calli, Hale Ögel Balaban, İhsan Yalçinkaya, and Umit Deniz Turan. 2010. The annotation scheme of the turkish discourse bank and an evaluation of inconsistent annotations. In *Proceedings of the fourth linguistic annotation workshop*, pages 282–289.

Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: A chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia.

Sandrine Zufferey and Liesbeth Degand. 2024. *Connectives and discourse relations*. Cambridge University Press.

6. Language Resource References

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2024.

Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0). Data/software, ÚFAL MFF UK, Prague, Czech Republic, LINDAT/CLARIAH-CZ: http://hdl.handle.net/11234/1-5813.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. 2006. Prague Dependency Treebank 2.0. *LDC2006T01*.

Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Eva Hajičová, Jiří Mírovský, Anna Nedoluzhko, Magdaléna Rysová, Veronika Pavlíková, Jana Zdeňková, Jiří Pergler, and Radek Ocelák. 2012. Prague Discourse Treebank 1.0. Data/software, ÚFAL MFF UK, Prague, Czech Republic, LINDAT/CLARIAH-CZ: http://hdl.handle.net/11858/00-097C-0000-0008-E130-A.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank version 3.0. *LDC2019T05*.

Magdaléna Rysová, Pavlína Jínová, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Jana Zdeňková, Veronika Scheller, and Šárka Zikánová. 2016. Prague Discourse Treebank 2.0. Data/software, ÚFAL MFF UK, Prague, Czech Republic, LINDAT/CLARIAH-CZ: http://hdl.handle.net/11234/1-1905.

Pavlína Synková, Jiří Mírovský, Marie Paclíková, Lucie Poláková, Magdaléna Rysová, Veronika Scheller, Jana Zdeňková, Šárka Zikánová, and Eva Hajičová. 2024. Prague Discourse Treebank 4.0. Data/software, ÚFAL MFF UK, Prague, Czech Republic, LINDAT/CLARIAH-CZ: http://hdl.handle.net/11234/1-5680.

Pavlína Synková, Magdaléna Rysová, Jiří Mírovský, Lucie Poláková, Veronika Scheller, Jana Zdeňková, Šárka Zikánová, and Eva Hajičová. 2022. Prague Discourse Treebank 3.0. Data/software, ÚFAL MFF UK, Prague, Czech Republic, LINDAT/CLARIAH-CZ: http://hdl.handle.net/11234/1-4875.

Šárka Zikánová, Pavlína Synková, and Jiří Mírovský. 2018. Enriched discourse annotation of PDiT subset 1.0 (PDiT-EDA 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), http://hdl.handle.net/11234/1-2906.