Announcing the Prague Discourse Treebank 3.0

Pavlína Synková, Jiří Mírovský, Lucie Poláková and Magdaléna Rysová
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic
{synkova, mirovsky, polakova, mrysova}@ufal.mff.cuni.cz

Abstract

We present the Prague Discourse Treebank 3.0 – a new version of the annotation of discourse relations marked by primary and secondary discourse connectives in the data of the Prague Dependency Treebank. Compared to the previous version (PDiT 2.0), the version 3.0 comes with three types of major updates: (i) it brings a largely revised annotation of discourse relations: pragmatic relations have been thoroughly reworked, many inconsistencies across all discourse types have been fixed and previously unclear cases marked in annotators' comments have been resolved, (ii) it achieves consistency with the Lexicon of Czech Discourse Connectives (CzeDLex), and (iii) it provides the data not only in its native format (Prague Markup Language, discourse relations annotated at the top of the dependency trees), but also in the Penn Discourse Treebank 3.0 format (plain text plus a stand-off discourse annotation) and sense taxonomy. PDiT 3.0 contains 21,662 discourse relations (plus 445 list relations) in 49 thousand sentences.

Keywords: discourse relations, pragmatic relations, Prague Discourse Treebank, Penn Discourse Treebank

1. Introduction

Corpora annotated manually with discourse phenomena are broadly considered to be important sources for both theoretical research on text coherence and for NLP tasks such as question answering, text summarization, machine reading comprehension or sentiment analysis. Automatic recognition of discourse relations either with present or absent connectives is often a core part of these tasks (an overview of the methods used for discourse parsing is given by Li et al., 2022, the most challenging task – recognition of discourse relations with absent connectives is surveyed by Xiang and Wang, 2023).

Discourse-annotated corpora have been created (using various theoretical frameworks) for many languages, incl. English (Prasad et al., 2008), Turkish (Zeyrek and Webber, 2008), Hindi (Oza et al., 2009), Italian (Tonelli et al., 2010), Arabic (Al-Saif and Markert, 2010), Dutch (Van Der Vliet et al., 2011), Tamil (Rachakonda and Sharma, 2011), French (Afantenos et al., 2012), Basque (Iruskieta et al., 2013) or Chinese (Zhou and Xue, 2015). However, reaching a good reliability in annotation of discourse relations seems to be a difficult task. This has been previously pointed out in a number of publications dedicated to this issue (Hoek and Scholman, 2017, Jínová et al., 2012a, Spooren and Degand, 2010) and can be further demonstrated by relatively low levels of inter-annotator agreement compared to manual annotations of other – less semantic – language phenomena (Mírovský and Hajičová, 2014), and high numbers of corrections reported in updated versions of discourse annotated corpora (Webber et al., 2019, Rysová et al., 2016).

As summarized by Hoek and Scholman (2017) from the study of Spooren and Degand (2010), the difficulty of discourse annotation comes from the fact "that coherence is a feature of the mental representation that readers form of a text, rather than of the linguistic material itself". Thus production of a large-scale corpus with high-quality discourse annotation is a resource-demanding task, often performed in stages, depending on available human and funding resources. High quality of annotation is reached gradually, by revisions of previously published and studied versions of the corpus.

The present paper introduces the Prague Discourse Treebank 3.0 (PDiT 3.0; (Synková et al., 2022)) – a new version of annotation of discourse relations marked by primary and secondary connectives in the data of the Prague Dependency Treebank (PDT; (Hajič et al., 2020)). Compared to the previous version (PDiT 2.0; (Rysová et al., 2016)[1]), PDiT 3.0 brings a largely revised annotation of discourse relations, reaches consistency with the lexicon of Czech discourse connectives CzeDLex and offers the data also in the Penn Discourse Treebank 3.0 (PDTB 3.0; (Prasad et al., 2019)) format and sense taxonomy.

In the following text, we first present the relevant data resources – the underlying PDT data, the discourse relation annotation in these data, and the lexicon of Czech discourse connectives CzeDLex (Section 2). Section 3 discusses the annotation revisions for version 3.0 in detail, Section 4 briefly describes the transformation to the PDTB 3.0 taxonomy and format. We conclude in Section 5.

[1] For a complete list of all previous versions, see https://ufal.mff.cuni.cz/pdit3.0.

LREC-COLING 2024, pages 1270–1279
20-25 May, 2024.

2. The Prague Discourse Treebank

The Prague Discourse Treebank (PDiT) is a layer of discourse relations annotated on the data of the Prague Dependency Treebank. Originally published separately (Hajič et al., 2006), the Prague Dependency Treebank (PDT) is now one of four treebanks published together as the Prague Dependency Treebank – Consolidated (Hajič et al., 2020), a corpus of Czech texts manually annotated on three layers of language description – (i) morphology, (ii) surface syntax and (iii) deep syntax (tectogrammatics).[2] The PDT itself contains 49,431 sentences (over 833 thousand tokens) annotated on all the layers.[3]

Discourse relations of PDiT are annotated on top of the deep syntax (tectogrammatical) layer of the PDT.

The tectogrammatical layer carries a complex annotation of a sentence following the Functional Generative Description proposed by Petr Sgall in the 1960s and later elaborated by him and his colleagues (Sgall et al., 1986). From the perspective of discourse annotation, it is important that on the tectogrammatical layer, (i) a sentence is represented as a dependency tree with nodes roughly corresponding to content words; the type of dependency relation between two nodes (such as *ACTor*, *PATient*, *CONDition* etc.) is kept in the *functor* attribute at the dependent node, (ii) the tree also contains nodes for elided entities (e.g. elided verbs), (iii) the left–right order of the nodes represents the information structure of the sentence (topic–focus articulation), and (iv) annotation of both grammatical and textual coreference is present.[4]

2.1. Discourse Annotation in PDiT

PDiT contains annotation of discourse relations explicitly signalled by discourse connectives, both primary (typically conjunctions or adverbs, such as *proto* [*therefore*] or *mezitím* [*meanwhile*]) and secondary (less fixed expressions such as *z toho důvodu* [*for that reason*] or *jinými slovy* [*in other words*]).[5] Each discourse relation connects two arguments which most often correspond to clauses (with a finite verb), compound sentences or, in some cases, to a sequence of sentences. Although some local hierarchies can be observed in the annotated discourse structures (Poláková et al., 2021), discourse annotation in PDiT is generally shallow and does not intend to form a hierarchical structure of a whole document.[6] Each relation is assigned a discourse type expressing its meaning (such as *reason–result*, *synchrony*, see Table 1). Additionally, PDiT contains annotation of lists (a separate type of relations),[7] annotation of headings, meta texts and genres (for each document one genre label, see Poláková et al., 2014).

The annotation of discourse relations in PDiT is inspired by the Penn Discourse Treebank (Prasad et al., 2008). While it follows the same basic principles, it also takes advantage of the Prague tradition of dependency treebanking and differs from the PDTB approach in at least four important aspects: (i) PDiT is annotated on top of deep-syntax dependency trees,[8] (ii) only discourse relations explicitly marked by connectives are annotated,[9] (iii) only relations with arguments containing finite verbs are annotated and (iv) PDiT uses its own repertoire of discourse relation types.

The original annotation proceeded in two consecutive steps – first, annotators went through all texts marking connectives and relations not represented on the tectogrammatical layer, second, relations represented on the tectogrammatical layer were semi-automatically transformed to the discourse layer (Jínová et al., 2012b).

The first version of PDiT (PDiT 1.0) only included primary connectives (conjunctions, adverbs, particles, some types of punctuation marks, some uses of pronouns and some types of idiomatic multi-word phrases), while candidates for other connective phrases were indicated just by annotators' comments; the second version (PDiT 2.0) reflected the division of connectives into primary and secondary; it included revisions of the relations from the previous version and new annotation of relations marked by secondary connectives. Version 2.0 contains approx. 21 thousand relations signalled by explicit connectives, out of which one thousand are relations with secondary connectives.

Technically, a discourse relation in PDiT is a connection between two tectogrammatical nodes which together with their subtrees represent the

[2] The Prague Dependency Treebank – Consolidated contains approx. 175 thousand sentences (2.7 million tokens) annotated on all the layers.

[3] More data are annotated only up to the surface syntax and even more only on the morphological layer.

[4] A dedicated study on the advantages of annotating discourse relations in tectogrammatical trees can be found in Mírovský et al. (2012).

[5] more about primary and secondary connectives in Rysová and Rysová (2014)

[6] unlike e.g. in the Rhetorical Structure Theory (RST; Mann and Thompson, 1988)

[7] Lists are not considered to be semantic relations. They merely structure the text as enumerations of items.

[8] This means that some intra-sentential discourse information could be extracted from the previous rich annotation of syntax (Jínová et al., 2012b).

[9] Implicit relations are annotated just in a sample of the data – approx. 2,600 sentences coming from 15 different genres. More information is available at https://ufal.mff.cuni.cz/pdit-eda1.0.

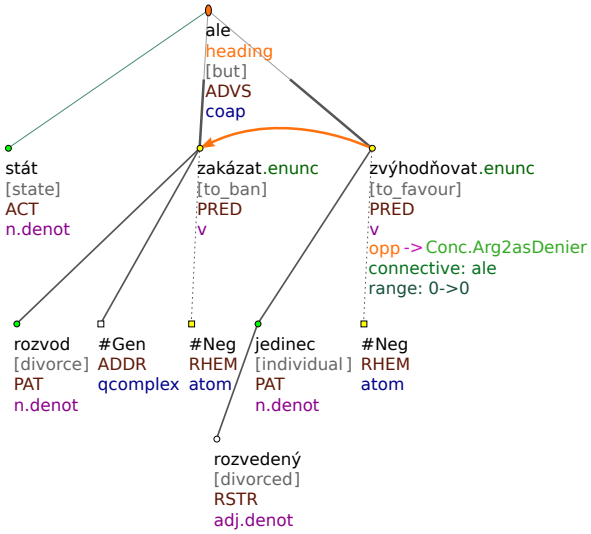

Figure 1: An intra-sentential discourse relation annotated in the tectogrammatical representation of the sentence from Example 1; the translations have been added here and are not a part of the published data.

two discourse arguments. This connection is visualized as an oriented orange arrow, see Figure 1 showing a tectogrammatical analysis with an annotated intra-sentential discourse relation for the sentence in Example 1.[10]

(1) *Stát nemůže rozvody zakázat,* ale **neměl by rozvedené jedince zvýhodňovat** (PDiT, ln94206_55)

[*The state cannot ban divorce* but **it should not favour divorced individuals**]

2.2. Lexicon of Czech Discourse Connectives

Lexicons of discourse connectives have recently become an integral part of the development of discourse corpora, either as a starting point for an annotation or as its possible outcome. For Czech, an online Lexicon of Czech Discourse Connectives – CzeDLex – is available[11] (Mírovský et al., 2017). It was developed on the basis of several discourse-annotated resources, mainly of the Prague Discourse Treebank's earlier versions.[12]

CzeDLex includes more than 200 level-one entries (e.g. *proto* [*therefore*], *však* [*however*]), most of them covering numerous complex forms (e.g. *a proto* [*and therefore*]) or modifications (e.g. *právě proto* [*exactly therefore*]). All entries are fully manually checked and provided with automatically extracted as well as manually added linguistic information (number of usages, usages with various discourse types, argument semantics, word order information, corpus examples, English translations etc.).

3. Annotation Revisions

The need to revise the discourse annotations published in the previous version of the Prague Discourse Treebank (PDiT 2.0) arose from three different sources:
(i) Already at the end of the original annotations, it was clear that some types of relations were more difficult to capture consistently than others. Namely, pragmatic relations (i.e. *pragmatic condition*, *pragmatic reason–result* and *pragmatic contrast*) and *explication* were defined relatively broadly, as they had not been established in the Czech syntactic tradition. As a result, individual annotators treated them somewhat differently.
(ii) The analysis of PDiT 2.0 data during the development of the CzeDLex lexicon revealed further inconsistencies in the annotation of individual connectives.
(iii) Finally, numerous annotators' comments in the original annotations often pointed to unclear cases that needed to be checked and revised.
The following subsections describe these three types of revisions for PDiT 3.0.

3.1. Pragmatic Relations

According to the PDiT annotation manual (Poláková et al., 2012), pragmatic relations were defined as (i) relations that involve some pragmatic phenomenon like subjectivity, complex inferencing, presuppositions etc., (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. The definition was deliberately broad: these phenomena had not been systematically described in the literature and authentic texts may show different types of relations of a complicated nature.
The analysis of all pragmatic relations in PDiT 2.0 (Poláková and Synková, 2021), carried out in the

[10] For examples in the paper, we follow the Penn Discourse Treebank convention of highlighting two arguments of a discourse relation 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.

[11] https://ufal.mff.cuni.cz/czedlex1.0 (Mírovský et al., 2021)

[12] As a supplementary resource, the Czech part of the Prague Czech–English Dependency Treebank (PCEDT-cz) was used (Hajič et al., 2012). PCEDT-cz is sentence-aligned Czech translation of all of the Penn Treebank-WSJ texts. CzeDLex entries were enriched by (manually checked) data from discourse annotation projected from the Penn Discourse Treebank 3.0 to the PCEDT-cz (Mírovský et al., 2021).

context of the research on global coherence patterns, confirmed that these relations (i) were not treated uniformly by the annotators, (ii) involved different types of difficulties in interpreting discourse relations, and (iii) some of these relations were erroneously annotated as pragmatic (e.g. due to lack of contextual knowledge). This analysis thus led to a complete re-annotation of all pragmatic relations in the PDiT data, leaving as pragmatic only those relations that involve some type of inference (either content-related knowledge or a speech act) as an argument or that are somehow stylistically inappropriate (so that we could not be sure how to interpret them). For example, the *opposition* relation in Example 2 can be understood because of a speech act – the author knows that the advice came too late and in a situation where anyone could give such an advice, but gives it anyway – which somehow weakens possible objections to the claim.

(2) *Po bitvě je každý generál.* **Kdybychom** však **hned od začátku přistoupili na osvědčený systém dražeb, mohli jsme si pár set milionů korun ve státní kase ušetřit.** (PDiT, ln94204_79)

[*After the battle, everyone is a general.* But, **if we had used a proven auction system right from the start, we could have saved a few hundred million crowns in the state treasury.**]

In PDiT 2.0, there were 17 *pragmatic condition* relations, 50 *pragmatic contrast* relations and 45 *pragmatic reason–result* relations. Approximately 30% of these relations were annotated as pragmatic due to lack of context knowledge – these contexts were re-annotated as *condition*, *reason–result* or *opposition* for the PDiT 3.0 data.

A small number of *pragmatic conditions* in the data indicated a possible high number of false negatives – i.e. cases where *pragmatic conditions* were annotated as *condition* due to insufficient annotator experience with these relations in the original annotation. Therefore, a probe was run to see if there were any *pragmatic conditions* annotated as *condition* in the whole data. As the probe confirmed this expectation, all 1,200 cases of condition were examined and about 90 *pragmatic conditions* were newly annotated – Example 3 illustrates one of them: the content of Arg2 (the if-clause) is not a condition for Arg1 – it is a factual event and the author claims the content of the first argument on the basis of this fact.

(3) Jestliže **český prezident a česká vláda ztratili nyní již definitivně trpělivost,** *neměl by se tomu Bonn podivovat a neměl by být zaskočen.* (PDiT, ln95045_068)

[If **the Czech president and the Czech government have now definitively lost their patience** *Bonn should not be surprised.*]

Moreover, some *pragmatic reason–result* relations were discovered due to re-annotation of *explication* relation in PDiT 2.0 (see below). So, in total, there are 106 *pragmatic conditions*, 29 *pragmatic contrasts* and 59 *pragmatic reason–results* in PDiT 3.0. An instance of *pragmatic reason–result* is given in Example 4. The pragmatic relation holds between the second argument and an inference of the type "this claim is clearly mistaken", because the client does not lose anything. The connective *přece* has no direct translation counterpart in English. Apart from causality, it often expresses a high degree of certainty of the speaker together with invoking the general validity of the given claim as in "as we all know".

(4) *Často se objevuje narážka, jak banka nahradí klientům vzniklou ztrátu.* **Klient** přece **o žádné své úspory nepřichází!** (PDiT, ln94211_9)

[*There is often an allusion to how the bank will compensate clients for the loss incurred.* After all, **the client does not lose any of his savings!**]

3.2. The *Explication* Relation

The relation of *explication* was established on the basis of Czech syntactic tradition and the presence of the connectives *totiž*, *vždyť*, *přece*, for which it is difficult to find exact English equivalents. They can be translated as *actually, you see, you know, as, after all*, or, in some contexts, they are implicitated (as in Examples 5 and 6). These connectives have a strong part of their meaning that can be translated as *you see*, *it is obvious*, *here I give my reasons for claiming that*. They are all ambiguous (they can signal many types of discourse relations – see CzeDLex), but they are often used when an explanation for the previous context is given. From a semantic point of view, this explanation is not necessarily given by means of a causal connection, but by means of a more elaborate reformulation of the left argument (the content of the arguments is synonymous, semantically close). This reformulation supports the claim in the left argument (Example 5) or helps understand its content (Example 6). In Example 5, the author supports a claim with details of how the malice was manifested; in Example 6, the author explains or justifies the phrase "any advantage".

(5) *Tyto provokace byly ovšem zjevně motivovány zlobou* : **všechny kocourkovské noviny si** totiž **do starosty s chutí rýply.** (PDiT, ln94207_72)

[*However, these provocations were clearly motivated by malice* : **all the Kocourkov newspapers took a dig at the mayor with gusto.**]

(6) *Takže s výjimkou běhu na 3 km překážek už nemají atleti-muži před atletkami žádný náskok.* **Ženy** totiž **skáčí i o tyči.** (PDiT, mf930713_105)

[*So, with the exception of the 3 km steeplechase, male athletes no longer have any advantage over female athletes.* **Women also do pole vaulting.**]

In the annotation manual (Poláková et al., 2012), the synonymy/similarity or closeness of the given propositions (i.e. a claim and an explanation for it) was established as the main criterion for distinguishing *reason–result* and *explication*.

The second criterion is the order of the arguments – while *reason–result* arguments can occur in either order (and often do in texts), claim and explanation do not normally allow the reverse order – the explanation follows the content being explained. A supporting criterion is the importance of the content of the argument – in the *explication* relation, the claim is more important than the following explanation, in contrast to the *reason–result* relation, where the major/minor importance of the content of the argument is not relevant.

Although the annotation instructions were clear, we expected that *reason–result* and *explication* would be difficult to distinguish in an authentic text. As this expectation was confirmed by a probe in the data, the whole set of relations annotated as *explication* in PDiT 2.0 was revised. Out of 279 relations in PDiT 2.0, only 147 remained as true *explications* in PDiT 3.0.

discourse type	PDiT 2.0	PDiT 3.0
COMPARISON		
concession	917	902
confrontation	665	686
correction	453	452
gradation	457	468
opposition	3,179	3,202
pragmatic contrast	50	29
restrictive opposition	271	285
CONTINGENCY		
condition	1,443	1,331
explication	279	147
pragmatic condition	17	106
pragmatic reason–result	45	59
purpose	419	421
reason–result	2,844	3,024
EXPANSION		
conjunction	7,712	7,746
conjunctive alternative	90	97
disjunctive alternative	271	271
equivalence	107	127
generalization	131	136
instantiation	157	208
specification	638	676
TEMPORAL		
precedence–succession	852	1,027
synchrony	226	262
Total	21,223	21,662

Table 1: Distributions of discourse types in PDiT 2.0 and PDiT 3.0

3.3. Expressions in Connective Use

The automatic process of connective extraction from the treebank to the CzeDLex lexicon lists for each potential discourse connective also corpus contexts with its non-connective usages (if present in the treebank texts). When going through these non-connective usages, we sometimes detected false negatives, i.e. contexts where a given expression actually was a connective and its annotation was omitted by mistake. Individual cases were collected, but expressions with more problematic contexts had to be systematically checked. Surprisingly, a group of 20 expressions with more problematic contexts included not only expressions with a low proportion of connective usages such as *stejně* [*equally, in the same way, also, still*] (12% of connective usages among all usages) or *ostatně* [*after all, for that matter*] (5% of connective usages), but also prototypical connectives such as the adverb *potom* [*then*] (85% of connective usages) or *navíc* [*above, moreover*] (78% of connective usages).

All occurrences of these expressions that were not associated with a discourse relation in the data were revised. In total, 1,400 instances of 20 expressions were revised, resulting in the annotation of approx. 360 new relations and modifications of 180 relations (mostly adding the given expression to the set of items representing the connective for the given relation).

3.4. Individual Contexts

In addition to systematically revising the annotations of many discourse-relevant expressions, individual contexts where corrections were needed were collected during the work on CzeDLex. In total, more than 300 individual contexts were collected and the necessary changes to the annotation were described. These changes included all aspects of discourse relations – their types,

connectives, argument spans – and were relevant both for primary and secondary connectives. In addition to individual contexts, systematic revisions were made: (i) for the temporal connectives *nato* [*after*], *dříve* [*earlier*], *později* [*later*], which in previous versions had been annotated rather randomly due to their adverbial nature, (ii) for the connective *s tím, že* [*along with* lit. *with that that*] with a rather vague meaning that seems to have confused annotators in the previous versions (see Examples 7 and 8, where this connective signals *reason–result* and *specification*, respectively), and (iii) for the connective *přece* [*after all*] with a strong presupposition that could lead to a (relevant) interpretation as an attitude marker rather than a discourse connective in the original version of the data.

(7) *Většinu výjezdů k závodům do zahraničí odmítal* s tím, že **má šichtu.** (PDiT, mf920922_075)

[*He refused most trips to competitions abroad*, saying [lit. with that that] **that he had a shift.**]

(8) *Navrhuje tvrdší postih recidivistů* , s tím, že **po třetím násilném činu by nekompromisně následovalo doživotí.** (PDiT, ln94209_18)

[*He proposes harsher punishment for recidivists*, saying [lit. with that that] **a third violent act would be uncompromisingly followed by life imprisonment.**]

In total, the examination of individual contexts led to the annotation of 140 new relations, 65 connectives were modified, 110 discourse types were changed and 15 relations were deleted.

3.5. Annotators' Comments

Corrections described in two previous sections showed that annotators' comments are a useful instrument for tracing problematic connectives and relations. Out of ca. 1,100 comments in PDiT 2.0, 240 were relevant for the present task.[13] The relevant comments often indicated mismatches between the manual and automatic part of connective detection. Example 9 shows a context where a discourse relation of *concession* is untypically signalled by a correlative connective: one part of the connective *i kdyby – přesto* [*even if – still*] occurs in the dependent clause (*i kdyby* [*even if*]), the second part in the main clause (*přesto* [*still*]). The annotator added a comment to the second part that it belonged to the *concession* relation, which was left for automatic processing because it could be easily obtained from the syntactic annotation on the tectogrammatical layer. However, the automatic procedure only took into account the first part (which is in a typical syntactic position), leaving the second part with no annotation.

(9) I kdyby **hnedka zítra řekla ČR, že smlouva je pasé**, přesto *by se teprve v březnu příštího roku mohla legislativně zbavit svých závazků vůči partnerovi z bývalé ČSFR.* (PDiT, cmpr9410_001)

[Even if **the Czech Republic were to tell tomorrow that the treaty is passé**, *it would* still *be able to legislatively get rid of its obligations to its partner from the former Czechoslovakia only in March of next year.*]

The second type of relevant comments indicated that the annotator wasn't sure whether a discourse relation existed in a given context, or that the context seemed to be of pragmatic nature. Individual contexts containing these comments were now manually reviewed and the annotation was revised in ca. 170 cases, including the annotation of ca. 40 new relations.

4. The PDTB 3.0 Taxonomy and Format in the PDiT 3.0 Data

In contrast to the previous versions, PDiT 3.0 includes also labels for all discourse relations in the PDTB 3.0 taxonomy (Webber et al., 2019), and the data are available both in the native PDT format (i.e., discourse relations annotated on top of deep-syntax dependency trees) and in the PDTB 3.0 format (plain text plus a stand-off discourse annotation). The transformation of PDiT types of discourse relations into PDTB 3.0 senses as well as the transformation into the PDTB 3.0 data format was described in detail in Mírovský et al. (2023). Here we only summarize the basic facts about the transformation and its results that are important for a stand-alone paper describing the PDiT 3.0 data.[14]

While the transformation of the data format was a technical rather than a theoretical task, the transformation of types into senses raised several theoretical questions – all of which are covered by Table 2, which is a slightly updated version of a similar table published in Mírovský et al. (2023).[15] As already discussed in Mírovský et al. (2023), most discourse types transform into a single sense

[13] the others were for example "error in the tectogrammatical tree"

[14] For more concise expressing in this section, we use the term "sense" for PDTB semantic relations and the term "type" for PDiT relations.

[15] Only second-level senses are listed in Table 2, as third-level senses are captured by the orientation of the discourse relation arrow in the PDiT data (e.g. for the *reason–result* type, the discourse arrow always begins in the argument where the reason is given, regardless of the relative position of the arguments in the text).

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

(e.g., *equivalence* into Expansion.Equivalence) and these instances could be transformed fully automatically.

Senses corresponding to more than one type (e.g. Comparison.Concession corresponds to *opposition* and *concession*) represent a loss of information and would only be an issue for transformation in the opposite direction (from senses to types). On the other hand, types that correspond to more than one sense (e.g. *condition* corresponds to Contingency.Condition and Contingency.Neg-condition) were studied in detail both in the annotation manuals and in the data, and were partially automatically transformed using various features (e.g. connectives, syntactic labels, presence of negation, argument order) that were found to be relevant for distinguishing these

sense	count
COMPARISON	
Concession.Arg1-as-denier	568
Concession.Arg2-as-denier	3,551
Concession+SA.Arg2-as-denier+SA	4
Contrast	780
Similarity	47
CONTINGENCY	
Cause+Belief.Reason+Belief	123
Cause+Belief.Result+Belief	7
Cause.Reason	1,750
Cause.Result	1,299
Cause+SA.Reason+SA	2
Cause+SA.Result+SA	4
Condition.Arg1-as-cond	48
Condition.Arg2-as-cond	1,237
Condition+SA	102
Neg-cause.NegResult	8
Neg-condition.Arg1-as-negCond	2
Neg-condition.Arg2-as-negCond	48
Purpose.Arg1-as-goal	6
Purpose.Arg2-as-goal	415
EXPANSION	
Conjunction	8,166
Disjunction	368
Equivalence	127
Exception.Arg1-as-excpt	6
Exception.Arg2-as-excpt	195
Instantiation.Arg1-as-instance	2
Instantiation.Arg2-as-instance	206
Level-of-detail.Arg1-as-detail	136
Level-of-detail.Arg2-as-detail	666
Substitution.Arg1-as-subst	61
Substitution.Arg2-as-subst	391
TEMPORAL	
Asynchronous.Precedence	686
Asynchronous.Succession	341
Synchronous	262
OTHER	48
Total	21,662

Table 3: Distributions of senses in PDiT 3.0

senses in the Czech data.[16] Pragmatic relations and *explication* were completely transformed manually, as there is no formal cue to distinguish cases with +Belief (B in Table 2), +SpeechAct (SA in Table 2) features from cases without them.
In total, approx. 42% of the 21.6 thousand PDiT relations were transformed into a single Penn sense;

[16] Details for all type–sense pairs can be found in Mírovský et al. (2023).

56% were transformed using rules based on linguistic features, and only about 2% of the relations had to be manually disambiguated in order to be transformed into a correct sense.
The distributions of types in PDiT 3.0 compared to PDiT 2.0 are given in Table 1, the distribution of senses in PDiT 3.0 is captured by Table 3 (a slightly updated version from Mírovský et al., 2023).[17]

5. Conclusion

This paper presents a new version of the Prague Discourse Treebank, PDiT 3.0. Compared to the previous version (PDiT 2.0), the annotation of discourse relations was thoroughly revised, correcting many inconsistencies revealed mostly by work on the Lexicon of Czech Discourse Connectives (CzeDLex) but by other sources as well. The new version of PDiT was published under the Creative Commons licence (Synková et al., 2022) both in its native format (Prague Markup Language) and in the Penn Discourse Treebank 3.0 format (Penn column format of discourse annotation accompanied by the original plain text).[18] All relations were also transformed to the Penn Discourse Treebank 3.0 taxonomy of senses, which makes the corpus compatible with this well established framework and easily accessible by the research community.
Annotation updates cover all annotated phenomena – presence of a relation, its type, connectives and arguments. The types of relations were systematically checked especially for all pragmatic relations and for *explication* relations, as these relations were difficult to handle in authentic text situations in the original annotation.
In total, about 3,600 contexts were checked manually, resulting in correction of about 850 relations, annotation of about 440 new relations, and deletion of about 50 relations.
PDiT 3.0 is consistent with the Lexicon of Czech Discourse Connectives (CzeDLex), meaning that connectives and their usages that appear in PDiT 3.0 are covered by the lexicon, providing additional information (incl. syntactic properties, corpus statistics and context examples).

[17] Line OTHER indicates a special type of *specification* relation that was omitted from the transformation because such a relation would not be annotated according to the PDTB annotation rules. These *specification* relations are part of the list relations in PDiT, i.e. they represent a relation connecting enumerated items (e.g. *1)*, *2)*) to a hypertheme (e.g. *the case has several problematic levels*) and, unlike "normal" *specifications*, are without a connective and/or also hold between nominal arguments.

[18] Thus, PDiT 3.0 can be opened and searched in both TrEd and Annotator tools (native annotation tools of the respective corpora).

6. Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project 22-03269S). 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 (LM2023062).

7. Ethics Statement

We honor the ethical code set out in the ACL Code of Ethics and there are no special ethical issues involved in this work.

8. 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 *LREC*, pages 2046–2053.

Jet Hoek and Merel Scholman. 2017. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (isa-13)*.

Mikel Iruskieta, Maria J Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The rst basque treebank: an online search interface to check rhetorical relations. In *4th workshop RST and discourse studies*, pages 40–49.

Pavlína Jínová, Jiří Mírovský, and Lucie Poláková. 2012a. Analyzing the most common errors in the discourse annotation of the prague dependency treebank. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 127–132, Lisboa, Portugal. Universidade de Lisboa, Edicoes Colibri, Lisboa.

Pavlína Jínová, Jiří Mírovský, and Lucie Poláková. 2012b. Semi-automatic annotation of intra-sentential discourse relations in PDT. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*, pages 43–58, Mumbai, India. Institute of Information Technologies (IIT) Bombay, Coling 2012 Organizing Committee.

Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022. A survey of discourse parsing. *Frontiers of Computer Science*, 16(5):165329.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Jiří Mírovský and Eva Hajičová. 2014. What can linguists learn from some simple statistics on annotated treebanks. In *Proceedings of 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 279–284, Tübingen, Germany. University of Tübingen, University of Tübingen.

Jiří Mírovský, Pavlína Jínová, and Lucie Poláková. 2012. Does tectogrammatics help the annotation of discourse? In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 853–862, Mumbai, India. IIT Bombay, Coling 2012 Organizing Committee.

Jiří Mírovský, Magdaléna Rysová, Pavlína Synková, and Lucie Poláková. 2023. Prague to penn discourse transformation. *The Prague Bulletin of Mathematical Linguistics*, (120):5–30.

Jiří Mírovský, Pavlína Synková, and Lucie Poláková. 2021. Extending coverage of a lexicon of discourse connectives using annotation projection. *The Prague Bulletin of Mathematical Linguistics*, (117):5–26.

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.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161.

Lucie Poláková, Pavlína Jínová, and Jiří Mírovský. 2014. Genres in the Prague Discourse Treebank. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1320–1326, Reykjavík, Iceland. European Language Resources Association.

Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, and Eva Hajičová. 2012. Manual for Annotation of Discourse Relations in Prague Dependency Treebank. Technical Report 47, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Lucie Poláková and Pavlína Synková. 2021. Pragmatické aspekty v popisu textové koherence. *Naše řeč*, 104(4):225–242.

Lucie Poláková, Jiří Mírovský, Šárka Zikánová, and Eva Hajičová. 2021. Discourse relations and connectives in higher text structure. *Dialogue and Discourse*, 12(2):1–37.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech. European Language Resources Association.

Ravi Teja Rachakonda and Dipti Misra Sharma. 2011. Creating an annotated tamil corpus as a discourse resource. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 119–123.

Magdaléna Rysová and Kateřina Rysová. 2014. The Centre and Periphery of Discourse Connectives. In *Proceedings of Pacific Asia Conference on Language, Information and Computing*, pages 452–459, Bangkok. Department of Linguistics, Faculty of Arts, Chulalongkorn University.

Magdaléna Rysová, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, and Šárka Zikánová. 2016. 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.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, Aravind K Joshi, et al. 2010. Annotation of discourse relations for conversational spoken dialogs. In *LREC*.

Nynke Van Der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. 2011. Building a discourse-annotated dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*, 35:108.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55(12):1–34.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for turkish: Annotating discourse connectives in the metu corpus. In *Proceedings of the 6th workshop on Asian language resources*.

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

9. Language Resource References

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.

Hajič, Jan and Bejček, Eduard and Bémová, Alevtina and Buráňová, Eva and Fučíková, Eva and Hajičová, Eva and Havelka, Jiří and Hlaváčová, Jaroslava and Homola, Petr and Ircing, Pavel and Kárník, Jiří and Kettnerová, Václava and Klyueva, Natalia and Kolářová, Veronika and Kučová, Lucie and Lopatková, Markéta and Mareček, David and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Novák, Michal and Pajas, Petr and Panevová, Jarmila and Peterek, Nino and Poláková, Lucie and Popel, Martin and Popelka, Jan and Rompoltl, Jan and Rysová, Magdaléna and Semecký, Jiří and Sgall, Petr and Spoustová, Johanka and Straka, Milan and Straňák, Pavel and Synková, Pavlína and Ševčíková, Magda and Šindlerová, Jana and Štěpánek, Jan and Štěpánková, Barbora and Toman, Josef and Urešová, Zdeňka and Vidová Hladká, Barbora and Zeman, Daniel and Zikánová, Šárka and Žabokrtský, Zdeněk. 2020. *Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University. [link].

Jan Hajič and Eva Hajičová and Jarmila Panevová and Petr Sgall and Silvie Cinková and Eva Fučíková and Marie Mikulová and Petr Pajas and Jan Popelka and Jiří Semecký and Jana Šindlerová and Jan Štěpánek and Josef Toman and Zdeňka Urešová and Zdeněk Žabokrtský. 2012. *Prague Czech-English Dependency Treebank 2.0*. University of Pennsylvania. Data/Software, Linguistic Data Consortium. [link].

Jiří Mírovský, Pavlína Synková, Lucie Poláková, Věra Kloudová, and Magdaléna Rysová. 2021. CzeDLex 1.0.

Rashmi Prasad and Bonnie Webber and Alan Lee and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. University of Pennsylvania. Data/Software, Linguistic Data Consortium. [link].

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.

Pavlína Synková and Magdaléna Rysová and Jiří Mírovský and Lucie Poláková and Veronika Sheller and Jana Zdeňková and Šárka Zikánová and Eva Hajičová. 2022. *Prague Discourse Treebank 3.0*. Institute of Formal and Applied Linguistics, Charles University. LINDAT/CLARIAH-CZ digital library. [link].