PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 120 APRIL 2023 5–30

Prague to Penn Discourse Transformation

Jiří Mírovský, Magdaléna Rysová, Pavlína Synková, Lucie Poláková

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

Abstract

The Prague and Penn styles of discourse annotation are close to each other in basic theoretical views and also in taxonomies of semantic types of discourse relations. A transformation from one of the annotation styles to the other should seemingly be a straightforward process. And yet, slight differences in the taxonomies and significant differences in the technical approaches present several interesting theoretical and practical challenges. The paper focuses on handling the most important issues in the transformation process from the Prague style to the Penn style of discourse annotation, in an effort to bring a valuable data resource – the Prague Discourse Treebank – closer to the international scientific community.

1. Introduction

Manually annotated text corpora have proven to be a multilateral and valuable resource for theoretical linguistic research, as well as for applied natural language processing (NLP), both as test data and for training machine-learning algorithms. The usefulness in the latter, however, has been multiplied in recent years with emergence of pre-trained deep learning methods and tools that use large unannotated data – raw texts – for training word embeddings (representation of (sub)words in a high-dimensional vector space) and for pre-training a deep neural network to "understand" basic language properties. Such a pre-trained system allows to fine-tune the model for a highly specific NLP task using a relatively small manually annotated data, leading to state-of-the-art results in many areas of NLP, as was first demonstrated with

Corresponding author: mirovsky@ufal.mff.cuni.cz

Cite as: Jiří Mírovský, Magdaléna Rysová, Pavlína Synková, Lucie Poláková. Prague to Penn Discourse Transformation. The Prague Bulletin of Mathematical Linguistics No. 120, 2023, pp. 5–30.
doi: 10.14712/00326585.025.

system BERT by Devlin et al. (2019).[1] Similar approach has since been successfully used for many other NLP tasks, including tasks closely related to text coherence, and specifically discourse relations.

The term *discourse relations* refers to semantic relations that connect two discourse units – segments of text expressing mostly individual events, states, situations (Zikánová et al., 2015). In Example 1, a discourse relation holds between two clauses and is signalled by an explicit discourse-structuring device, a connective *but*.

(1) *Profit may be low*, but **at least costs should be covered**. (PDTB, wsj_0051)
[*Zisk může být malý*, ale **měly by se alespoň zaplatit náklady**.[2] (PCEDT, wsj_0051)]

Depending on a chosen taxonomy, a discourse relation can be classified in one of (usually several tens of) semantic types (e.g., in Example 1, *Comparison.Concession.Arg2-as-denier*, or in another taxonomy, *opposition*). 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*.

A growing interest in text coherence-aware methods can be traced in many areas of natural language processing, including tasks such as machine translation (Xiong et al., 2019; Meyer and Webber, 2013), text generation (Kiddon et al., 2016), summarization (Zhang, 2011), information extraction, opinion mining (Turney and Littman, 2003), coherence evaluation (Rysová et al., 2016), or machine translation evaluation (Bojar et al., 2018). Many of these tasks incorporate a discourse parser in the text pre-processing and, of course, discourse parsing methods have received a lot of attention from the NLP community, including two CoNLL shared tasks (Xue et al., 2015, 2016). Recently, pre-trained deep learning systems such as BERT have spread also to this field: Shi and Demberg (2019) use BERT for classification of so-called implicit discourse relations, outperforming the state of the art. Similarly, Mírovský and Poláková (2021) show that information about the presence of a discourse connective can be incorporated into the BERT framework and that text corpora annotated manually with explicit discourse relations can be successfully used to fine-tune BERT to classify also explicit discourse relations (both in Czech and English).

Several theoretical frameworks for discourse relations representation were developed and used both for theoretical description and for corpora annotation in last decades, with two of them being probably most influential: the approach developed and first used for the annotation of the Penn Discourse Treebank (PDTB; Prasad et al.,

[1] The authors used BERT to reach or improve state-of-the-art results for tasks such as language understanding, question answering and language generation.

[2] 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.

2008; Prasad et al., 2019), and the Rhetorical Structure Theory (RST; Mann and Thompson, 1988; Taboada and Mann, 2006). While the PDTB model works "locally", i.e. it looks for discourse relations between two (mostly) adjacent clauses or sentences, the RST represents a "global" coherence model, considering each document as a whole to be hierarchically interconnected by rhetorical relations, forming a single tree-like structure.

The Prague Discourse Treebank (PDiT, Poláková et al., 2013; Rysová et al., 2016) is a large corpus of Czech newspaper texts manually annotated with discourse relations. The annotation of discourse relations in PDiT adopts the "local" approach to discourse relations representation and in many aspects is similar to the PDTB approach and is inspired by it (see Section 2). In fact, the relative theory-neutrality of the PDTB approach, the easy applicability of its annotation scheme also to languages other than English, a usually fair inter-annotator agreement and – given its relative simplicity – the possibility to manually annotate a relatively large text corpus, attracted many followers and has been employed in numerous annotation projects.[3] Also both CoNLL shared tasks mentioned above used data annotated according to the PDTB principles.

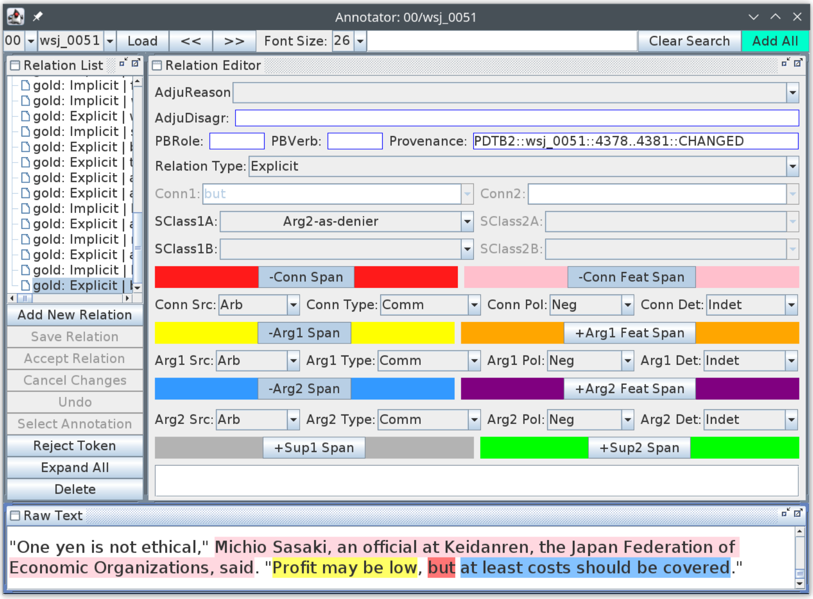
However, in the Prague Discourse Treebank, unlike most other discourse-annotated corpora, the annotation was not done on top of raw texts but instead on dependency trees of a deep-syntactic layer called *tectogrammatics*. It brings numerous advantages (resolved ellipses, arguments corresponding to subtrees, some relations already captured in the syntax tree, see Mírovský et al. (2012) for details). Yet, a substantial complexity of the native data format of PDiT presents a serious hindrance for any researcher not familiar with the data format and with the annotation theory of the deep-syntactic (tectogrammatical) layer of the corpus.

The present paper deals with theoretical and practical issues of the transformation of the discourse relations annotation of the Prague Discourse Treebank from its original (Prague) format and formalism to the Penn Discourse Treebank framework. In Section 2, we briefly describe the two involved discourse annotation frameworks – the *Penn style* and the *Prague style*. In Section 3, we describe in detail transformation steps from the Prague taxonomy of semantic types (called *discourse types*) to the Penn taxonomy of semantic types (called *senses*). In Section 4, we evaluate the results of the transformation and discuss main differences in sense distributions in the transformed PDiT vs. the PDTB. We conclude and outline future directions in Section 5.

2. Prague and Penn Styles of Discourse Annotation

This section shortly describes relevant parts of the two discourse annotation frameworks under consideration, i.e. the Penn style used in the Penn Discourse Treebank (PDTB), and the Prague style used in the Prague Discourse Treebank (PDiT). We start

[3] Prasad et al. (2008, 2019) (English), Oza et al. (2009) (Hindi), Zeyrek and Kurfalı (2017) (Turkish), Danlos et al. (2012) (French), Zhou and Xue (2012) (Chinese), and many others.



AdjuReason:
AdjuDisagr:
PBRole: PBVerb: Provenance: PDTB2::wsj_0051::4378..4381::CHANGED
Relation Type: Explicit
Conn1: but Conn2:
SClass1A: Arg2-as-denier SClass2A:
SClass1B: SClass2B:
Conn Src: Arb Conn Type: Comm Conn Pol: Neg Conn Det: Indet
Arg1 Src: Arb Arg1 Type: Comm Arg1 Pol: Neg Arg1 Det: Indet
Arg2 Src: Arb Arg2 Type: Comm Arg2 Pol: Neg Arg2 Det: Indet

"One yen is not ethical," Michio Sasaki, an official at Keidanren, the Japan Federation of Economic Organizations, said. "Profit may be low, but at least costs should be covered."

Figure 1. Annotation of the sentence from Example 1 in the PDTB annotation tool Annotator

with the Penn style and follow with the Prague style and its main differences from the Penn style. To easily distinguish the two taxonomies of semantic types in the subsequent text, we use the term *sense* for a semantic type in the Penn style, and the term *discourse type* for a semantic type in the Prague style of discourse annotation.

2.1. Penn Style of Discourse Annotation

The Penn style of discourse annotation employed in the PDTB follows a lexically-grounded approach to annotation of discourse relations (Webber et al., 2003): A discourse connective is a lexical anchor of a discourse relation that holds between two text spans called arguments. The annotation follows the *minimality principle*: the extent of the arguments is marked only as large as needed to interpret the discourse relation properly.

The connective signals the sense of the discourse relation; if it is absent, the relation is called *implicit*. The sense taxonomy is organized into three levels, with four major classes on the first level and 35 detailed senses[4] on the third level, which also reflects the asymmetry of some of the senses. Table 2 (see Section 4 below) lists all senses for explicit discourse relations in the PDTB 3.0.

If applicable, discourse relations can carry additional information: (i) a *second sense*, if it is distinctly present in the relation beside the first, most prominent sense, (ii) an *attribution* of the relation and of the arguments (i.e., parts of the text that indicate the authors of the statements represented by the relation/arguments), and (iii) a *supplement*, i.e. additional pieces of text beyond the minimality principle that play a supplementary role in interpreting the discourse relation.

In the PDTB 3.0, discourse relations are marked in a stand-off way on top of plain texts (i.e., no text pre-processing needed), and the two arguments, the connective (if present) and other properties are delimited using links to the plain text, i.e. as text spans. In total, there are approx. 25 thousand explicit discourse relations annotated in the PDTB 3.0.

Figure 1 shows the annotation of the discourse relation in the sentence from Example 1 in the PDTB 3.0, displayed in the PDTB annotation tool Annotator (for details on the tool, see Lee et al., 2016).

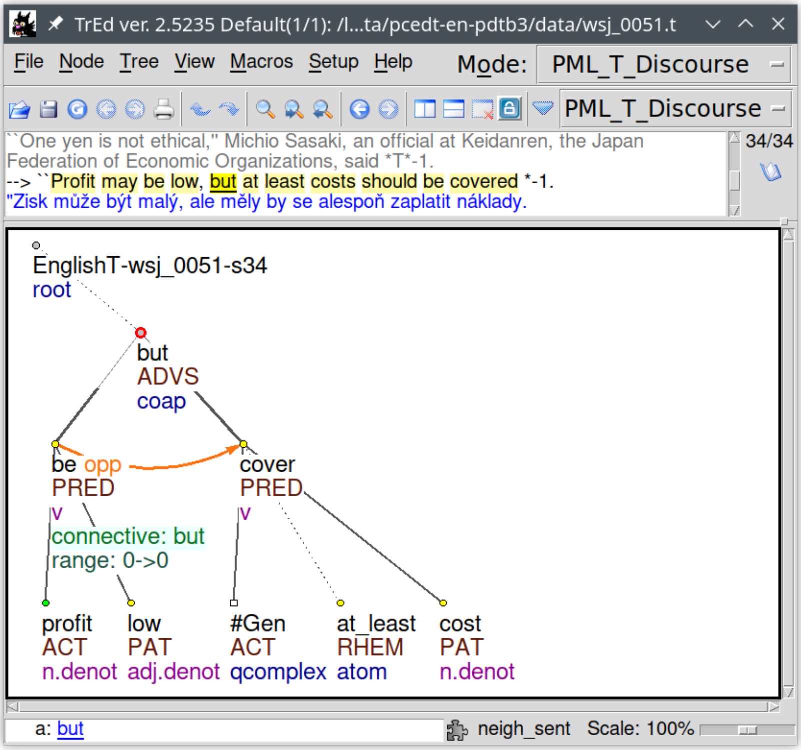
2.2. Prague Style of Discourse Annotation

Annotation of discourse relations in Czech was to a great extent inspired by the PDTB approach (Poláková et al., 2013). The Prague style of discourse annotation follows the Penn style in marking discourse connectives, their two arguments and the relation semantics, and it also follows the minimality priciple. The list of semantic types of discourse relations (*discourse types*) is close to the list of senses used in the PDTB (especially to the PDTB 3.0 hierarchy), slightly adapted according to the Czech syntactic tradition.[5] The Czech tradition of dependency treebanking was embraced also by incorporating the discourse annotation into the stratificational system of a multi-layered language description. Discourse relations thus have not been annotated on plain texts but instead on top of the deep-syntactic (tectogrammatical) layer of the underlying corpus, the Prague Dependency Treebank (PDT; its most recent version was published as a part of the Prague Dependency Treebank - Consolidated 1.0, Hajič et al., 2020).

The underlying corpus, the PDT, is a richly annotated language resource with a multi-layer annotation architecture: (i) a word layer (w-layer), where the plain text is segmented into documents and paragraphs and tokenized, (ii) a morphological layer (m-layer) with segmentation to sentences, all tokens get a lemma and a morphological

[4] 35 is the number of different senses actually appearing in the PDTB 3.0 incl. +*Belief* and +*SpeechAct* aspects.

[5] There is e.g. a *gradation* relation in the Prague taxonomy, prototypically expressed by multi-part *not only... but also* connective).



``One yen is not ethical," Michio Sasaki, an official at Keidanren, the Japan Federation of Economic Organizations, said *T*-1.
--> ``Profit may be low, but at least costs should be covered *-1.
"Zisk může být malý, ale měly by se alespoň zaplatit náklady.

Figure 2. Annotation of the sentence from Example 1 in the Prague discourse annotation tool TrEd

tag, (iii) a surface-syntactic layer (analytical, a-layer): a dependency tree capturing surface syntactic relations such as subject, object, adverbial, (iv) a deep-syntactic layer (tectogrammatical, t-layer): a dependency tree capturing deep syntactic relations (semantically interpreted using labels called functors), ellipses, valency and coreference.

Two major versions of the annotation of discourse relations in the PDT data were published as the Prague Discourse Treebank 1.0 and the Prague Discourse Treebank 2.0. The first version (PDiT 1.0) captured discourse relations marked by explicit connectives (covering conjunctions, adverbs, particles, some types of punctuation marks, some uses of relative pronouns and some types of idiomatic multi-word phrases) and arguments (text units) they connect (Poláková et al., 2013; Mírovský et al., 2014; Zikánová et al., 2015). The data were later updated in PDiT 2.0 (Rysová et al., 2016) with annotation reflecting the division of connectives into *primary* connectives (grammaticalized single-word units or non-compositional multi-word units) and *secondary* connectives[6] (not yet fully grammaticalized, compositional structures such as *this is the reason why*, *under these conditions*, etc.; Rysová and Rysová, 2014, 2018). In total, there are 21 thousand annotated occurrences of discourse relations expressed by explicit connectives, out of which 20 thousand are expressed by primary connectives.

In contrast with the Penn style, the Prague Discourse Treebank annotation does not include implicit relations, second senses of relations (i.e., always a single sense is attached to a relation), and also attribution is not annotated.[7]

Figure 2 shows the annotation of the discourse relation in the sentence from Example 1 in the Prague style of discourse annotation,[8] displayed in tree editor TrEd[9] (Pajas and Štěpánek, 2008). The discourse relation is expressed by an arrow connecting roots of subtrees corresponding to the arguments of the relation. Its direction indicates the argument semantics (i.e., it corresponds to the third level of senses in the Penn style).

The upcoming Prague Discourse Treebank 3.0 brings a substantial revision of discourse types assignment from the previous release, based in large part on the prior work on the Lexicon of Czech Discourse Connectives (CzeDLex; Mírovský et al., 2021) and, as elaborated and discussed in the rest of the present paper, it offers the annotation of discourse relations also in the Penn style (incl. the Penn sense taxonomy).

3. Transformation of Senses

The transformation process from the Prague style to the Penn style of discourse annotation consists of two separate parts: (i) transformation of the data format, which – although complex – is more a technical than a theoretical problem and we mention it only briefly in Section 3.7, and (ii) transformation of Prague discourse types to Penn senses. The latter brings up a number of theoretical questions that are discussed in the subsequent text.

Table 1 shows a transformation table from Prague discourse types (on the left) to the second level of Penn senses (on the right), based on a detailed study of the annotation manuals and the data of the two corpora. For asymmetric relations, the third level of senses (the argument semantics) is assessed from the direction of the discourse arrow in the Prague annotation.

At a first glance we can make several observations: (i) most discourse types transform to a single sense, (ii) some discourse types transform to two senses, (iii) some

[6] roughly corresponding to *alternative lexicalizations* in the Penn style

[7] More complete discourse annotation, incl. annotation of implicit relations, has been done on a relatively small part of the PDiT data only and published separately as Enriched Discourse Annotation of PDiT Subset 1.0 (PDiT-EDA 1.0; Zikánová et al., 2018).

[8] The underlying tectogrammatical tree comes from the Prague Czech–English Dependency Treebank (PCEDT; Hajič et al., 2012).

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

senses correspond to more than one discourse type, and (iv) the division to the four major classes[10] sometimes changes after the transformation.

Observations (iii) and (iv) are not substantial for the present task of Prague to Penn transformation. The Penn senses that correspond to more than one Prague discourse type (e.g., *Comparison.Concession*) merely account in this transformation direction for an (unavoidable) information loss and would only represent an issue for the opposite direction of transformation (Penn to Prague).[11]

Changes in the division to the four main sense classes are a matter of different underlying theoretical categorizations. They take place in such cases of Prague discourse types that were newly introduced for the Prague annotation and did not exist in the PDTB 2.0. The *restrictive opposition* discourse type, for instance, is a wider relation than *Expansion.Exception*, it also encompasses a more relaxed restriction of the content of the other argument. This includes a contrastive (or polarity-change) feature and also contrastive connectives are often used. The affiliation of *correction* and *gradation*, the other two Prague-only labels, to the Comparison class is based on the same principle, cmp. for example the contrastive feature in the complex *not only but also* connective.

On the other hand, observations (i) and (ii) are of the utmost importance. Discourse types that transform to a single sense can be processed without further consideration and represent a fully automatic part of the transformation. Discourse types that transform to two senses need further attention.

Our effort was aimed at discovering to which extent these ambiguous discourse types can be processed automatically with a satisfying success rate and which part of the data needs to be processed manually. The rest of this section is dedicated to a thorough analysis of transformation needs of the individual ambiguous discourse types.

3.1. *Comparison.Similarity* from *conjunction*

One of the relations that is present in the PDTB taxonomy but not in that of PDiT is a relation of *Comparison.Similarity*. *Comparison.Similarity* in the PDTB (Webber et al., 2019) is characterized as follows: “This tag is used when one or more similarities between Arg1 and Arg2 are highlighted with respect to what each argument predicates as a whole or to some entities it mentions.”

This sense in PDiT was captured under the relation of *conjunction*. In the preparation of the transformation process, we examined all PDTB occurrences of *Comparison.Similarity* relation and took under scrutiny all connectives used for this sense.

[10] *TEMPORAL, CONTINGENCY, COMPARISON, EXPANSION*

[11] Regarding the sense *Expansion.Level-of-detail* corresponding to two Prague discourse types (*specification* and *generalization*), this ambiguity in the opposite transformation process would easily be solved by taking into account the third level of the PDTB sense hierarchy (argument ordering) and the direction of the relation in the Prague style.

PDiT discourse type	PDTB 3.0 sense(s)
TEMPORAL	
precedence-succession	Temporal.Asynchronous
synchrony	Temporal.Synchronous
CONTINGENCY	
reason-result	Contingency.Cause, Contingency.Negative-cause
pragmatic reason-result	Contingency.Cause+Belief, Contingency.Cause+SpeechAct
condition	Contingency.Condition, Contingency.Negative-condition
pragmatic condition	Contingency.Condition+SpeechAct, Contingency.Negative-condition+SpeechAct
purpose	Contingency.Purpose
explication	Contingency.Cause+Belief
COMPARISON	
confrontation	Comparison.Contrast
opposition	Comparison.Concession
pragmatic contrast	Comparison.Concession+Belief, Comparison.Concession+SpeechAct
restrictive opposition	Expansion.Exception, Comparison.Contrast
concession	Comparison.Concession
correction	Expansion.Substitution
gradation	Expansion.Conjunction
EXPANSION	
conjunction	Expansion.Conjunction, Comparison.Similarity
instantiation	Expansion.Instantiation
specification	Expansion.Level-of-detail
generalization	Expansion.Level-of-detail
equivalence	Expansion.Equivalence
conjunctive alternative	Expansion.Disjunction
disjunctive alternative	Expansion.Disjunction

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

Then we looked for their counterparts in Czech that occurred within the PDiT relation of *conjunction*. In this way, we found *Comparison.Similarity* connectives in Czech, namely single-word connectives *obdobně* [*similarly*] and *podobně* [*similarly*], and their complex variants such as *podobně i* [*similarly also*] or *podobně jako* [*similarly as*], as well as complex connectives containing the word *stejně* [*equally, still*] such as *stejně tak* or *stejně jako* [both meaning *likewise*], see Example 2.

(2) *Také v tomto případě jde o autonomní aktivitu finanční instituce, nota bene na vládě nezávislé, zmocněné k tomu zákonem.* Podobně **vláda využívá mzdové regulace, vyžaduje-li to nárůst inflace.** (PDiT, ln94200_126)

[*Also in this case, it is an autonomous activity of a financial institution, nota bene independent of the government, authorized to do so by law.* Similarly, **the government uses wage regulation if inflation increases.**]

The one-word connective *stejně* [*equally, still, anyway*] did not appear expressing the relation of *Comparison.Similarity*, and therefore it was not covered within this sense, see Example 3.

(3) *Demokracii si můžeme dovolit, protože máme nejlepší a historicky spravedlivý program a národ nás miluje.* **O moc** stejně **nepřijdeme, protože volby vyhrajeme.** (PDiT, ln95048_117)

[*We can afford democracy because we have the best and historically just program and the nation loves us.* **We won't lose power** anyway, **because we will win the elections.**]

3.2. *Contingency.Negative-condition* from *condition*

Another relation that required a deep analysis was the relation of *Contingency.Negative-condition*. In the PDTB manual (Webber et al., 2019), this relation is defined as follows: "This tag is used when one argument (the antecedent) describes a situation presented as unrealized, which if it doesn't occur, would lead to the situation described by the other argument (the consequent). There are distinct senses for interpreting the arguments in terms of semantics or speech acts, with the default being semantics. The label *Contingency.Negative-condition.Arg1-as-negCond* is used when Arg1 describes the antecedent and Arg2, the consequent."

In the analysis of *Contingency.Negative-condition* annotated for English, we focused especially on specific connectives used for this relation and we searched for their counterparts in Czech. We found the following connectives in Czech that were originally annotated as a pure *condition* in PDiT: *jinak* [a counterpart of English *otherwise* and *lest*], *nebo* or *buď_nebo* [counterparts of English *or* and *either_or*] and *aniž* [a counterpart of English constructions containing *without*].

The most challenging situation appeared to be with the connective *unless* (the most frequent connective for *Contingency.Negative-condition* in the PDTB). Czech language does not have a direct counterpart for this English connective. Thus we faced a complicated issue of how to find Czech contexts in PDiT that correspond meaningfully to English contexts with the connective *unless*.

The connective *unless* contains negation in its sense, but it does not simply mean "if not". However, the presence of negation in the Czech sentence was a basic condition for the search of Czech counterparts of English sentences with *unless*.

The reliable cases that could be marked as *Contingency.Negative-condition* automatically were those in which a connective expressing discourse type *condition* (*pokud, když, -li* [all meaning *if*]) and a connective such as *tedy* [*that is*], *ovšem* or *však* [both meaning *however*] occurred together in the sentence containing a negation, see Example 4.

(4) *Za rok tu jsem znova*, tedy pokud **mě nepřejede auto.** (PDiT, ln94207_54)
[*I'll be here again in a year* unless **I get run over by a car.**]

However, the second connective (like *tedy* [*that is*] in the example) occurs explicitly in the sentence rather rarely. Therefore, we were looking for other tendencies that characterize the relation of *Contingency.Negative-condition* in Czech.

It turned out that these are the order of the discourse arguments in combination with a particular connective. A big portion of cases that were evaluated as *Contingency.Negative-condition* contained a connective *pokud* or *-li* [both meaning *if*] in the second argument, see Examples 5 and 6.

(5) *Celý rok jsme přečkali bez změny ceny,* **nepočítáme**-li **zvýšení v souvislosti se zařazením barevného televizního magazínu Duha jako přílohy LN**. (PDiT, ln94210_111)
[*We went the whole year without a price change* unless **we count the increase in connection with the inclusion of the color TV magazine Duha as a supplement to LN**.]

(6) *Mělo by to stačit*, pokud **se nevynoří něco nenadálého**. (PDiT, ln94205_130)
[*That should be enough* unless **something unexpected comes up**.]

3.3. *Contingency.Negative-cause* from *reason–result*

Special attention also had to be paid to the relation of *Contingency.Negative-cause.negResult*. According to the PDTB 3.0 manual, this relation "is used when Arg1 gives the reason, explanation or justification that prevents the effect mentioned in Arg2." It also mentions that the relation "was specifically introduced for the lexico-syntactic construction 'too X to Y'."

This construction corresponds to Czech complex connectives *na to, aby* or *k tomu, aby* that occur together with an adjunct expressing manner by specifying extent or intensity of the event or a circumstance, such as *příliš* [*too* (*much*)], see Example 7.

(7) *Jsem* příliš *mladý* na to, abych **žil se založenýma rukama**. (PDiT, mf920925_120)
[*I'm* too *young* to **live with folded hands**.]

These cases were annotated as a relation of *reason–result* in PDiT. However, all of them have been provided with a comment by an annotator that these constructions are rather specific and require further attention. In this regard, the annotation of these cases as *Contingency.Negative-cause.negResult* provides an effective solution also for Czech.

All these cases have a dependent clause labelled on the underlying tectogrammatical layer by the AIM functor[12] and these cases were a part of discourse annotation. To be sure that all such constructions were treated the same way, we searched for them also in compound sentences with a dependent clause labelled with RESL functor,[13] which was originally omitted from the discourse annotation, because a vast majority of RESL clauses do not have a discourse interpretation. In this way, three additional cases were found to be interpreted as *Contingency.Negative-cause.negResult* (and *reason–result* in the Prague taxonomy).

3.4. *Comparison.Contrast* from *restrictive opposition*

Another issue to be solved concerned the relation of *restrictive opposition*. *Restrictive opposition* in the Prague style is a relation in which the validity of the first argument is limited by the content of the second argument or the second argument expresses an exception to the first one (see the PDiT annotation manual, Poláková et al., 2012). So, the scope of the relation is wider than the one of the *Expansion.Exception* PDTB sense.

We primarily converted Prague relations of *restrictive opposition* to the PDTB 3.0 *Expansion.Exception*[14] but sometimes also to *Comparison.Contrast*.[15] We assumed the relation of *Comparison.Contrast* in cases where *restrictive opposition* was not accompanied by the use of a functor RESTR[16] on the underlying tectogrammatical layer.

Firstly, we manually evaluated cases of intra-sentential relations of *restrictive opposition* in a complex sentence in which the subordinate clause did not contain the

[12] This label is used for non-obligatory modifications that express purpose, the intended result or the aim (Mikulová et al., 2005).

[13] This label is used for a non-obligatory modification that "expresses manner by specifying the result of the event" (Mikulová et al., 2005).

[14] "This tag is used when one argument evokes a set of circumstances in which the described situation holds, and the other argument indicates one or more instances where it doesn't," see the PDTB 3.0 manual (Webber et al., 2019).

[15] "Contrast is used when at least two differences between Arg1 and Arg2 are highlighted," see the PDTB 3.0 manual (Webber et al., 2019).

[16] Label RESTR (restriction) is used for a non-obligatory modification that "expresses manner by specifying an exception/restriction" (Mikulová et al., 2005).

functor RESTR. We found out that the most cases of *Comparison.Contrast* appeared in sentences with connectives *však* [*however*] and (*i*) *když* [*although*].

In the next step, we thus limited our analysis to these connectives and extended the search also to inter-sentential relations. We found altogether 114 occurrences of such a type of sentence and manually marked 86 of them as a relation of *Comparison.Contrast*, see Example 8.

(8) *Lidé na všech stupních řízení jsou schopní, mají snahu se dále učit.* **Chybí jim** však **zkušenosti z dlouhodobého působení.** (PDiT, cmpr9410_010)

[*People at all levels of management are efficient and eager to learn.* However, **they lack long-term experience**.]

The rest of these sentences were annotated as *Expansion.Exception*, see Example 9.

(9) *Jeho návrh hovoří o šecích, které by následně získaly domácnosti od státu na placení všech faktur za energie, které domácnost využije.* **Vyloučeny by** však **byly motorové kapalné pohonné hmoty**. (PDiT, cmpr9410_049)

[*His proposal talks about checks that households would subsequently receive from the state to pay all invoices for energy that the household uses.* However, **liquid motor fuels would be excluded**.]

3.5. Pragmatic Relations

Three pragmatic relations were established in the Prague taxonomy of discourse types – namely *pragmatic reason–result*, *pragmatic condition* and *pragmatic contrast*. Although these relations were originally inspired by the PDTB 2.0 pragmatic relations, they were in the Prague style defined broader: these labels were used for cases where the semantics and the form do not correspond to each other. In a vast majority of cases, such a relation holds between one argument and a content that is inferred from the other argument. Analysis of all pragmatic relations in PDiT (Poláková and Synková, 2021) showed that this discrepancy/inference can be of various kinds, two of them corresponding to PDTB 3.0 relations with +*Belief* and +*SpeechAct* aspects (namely *Contingency.Cause+Belief*, *Contingency.Cause+SpeechAct*, *Contingency.Condition+SpeechAct* and *Comparison.Concession+SpeechAct*). *Contingency.Cause+Belief* "is used when evidence is provided to cause the hearer to believe a claim. The belief is implicit." (PDTB 3.0 manual; Webber et al., 2019), tags with +*SpeechAct* aspect were used when a relation holds between an argument and an implicit speech act represented by the other argument (PDTB 3.0 manual) – see Example 10.

(10) Jestliže **chcete slyšet můj postoj k rozhodnutí poroty**, *je to neslýchaný projev neúcty k práci druhého.* (PDiT, lnd94103_102)

[If **you want to hear my take on the jury's decision**, *it's an unheard of disrespect for someone else's work.*]

In contrast to the Penn definition, *pragmatic reason–result* relations in the Prague style corresponding to the Penn relation of *Contingency.Cause+Belief* have also the subjectivity aspect – a claim or provided evidence was a highly subjective one, as showed by Example 11.

(11) *Nemají se za co omlouvat, ale zároveň se nesmějí starat jen o sebe a svá konta.* Proto **by měli deset procent z vyhraných peněz věnovat na charitu**. (PDiT, ln94208_106)

[*They have nothing to apologize for, but at the same time they must not only care about themselves and their accounts.* Therefore, **they should donate ten percent of the money won to charity**.]

Besides these relations (corresponding to Penn +*Belief* and +*SpeechAct* relations), there were also cases where pragmatic relations in PDiT were annotated because of a complicated inference resulting from a cultural context, and cases with broken coherence caused by a formulation clumsiness. These relations were transformed to Penn senses without the +*Belief* and +*SpeechAct* aspects.

Discourse types of all pragmatic relations in PDiT were transformed to the corresponding Penn senses manually because there is no formal clue for distinguishing cases with +*Belief* and +*SpeechAct* aspects, and cases without them.

Altogether, 35 of 100 pragmatic relations in PDiT were transformed to relations with +*Belief* or +*SpeechAct* aspects, leaving the rest of them labelled as *Contingency.Cause*, *Contingency.Condition* or *Comparison.Concession*.

The above analysis has shown that the relation of *pragmatic condition* in PDiT was annotated quite rarely, implying a possible high number of false negatives. So a probe was performed in the whole data to see if some *pragmatic conditions* were by mistake annotated as *conditions*. As some *pragmatic conditions* were indeed found in the analyzed sample of relations of *condition*, all *condition* relations were then checked manually and 92 *pragmatic conditions* (corresponding to *Contingency.Condition+SpeechAct*) were newly annotated. One of them is given in Example 12.

(12) Kdybych **měl jmenovat konkrétní autory**, *byla by jich spousta*. (PDiT, ln95048_050)

[If **I should name specific authors**, *there would be lots of them.*]

3.6. *Specification* with the List Relation

The Prague annotation style recognizes a special type of relation called *list*. The list relation holds between enumerated items (i.e. *first, second*; *1*), *2*) etc.) and these items as a whole are connected with its hypertheme (i.e., sentences such as *there are several problematic issues*) by a *specification* relation that can (contrary to *specification* relation not related to a list) be without a connective or can hold between nominal arguments.

As the list relation does not have a counterpart in the Penn style of annotation,[17] it is omitted from the transformation. However, the introductory *specification* relation has its counterpart in *Expansion.Level-of-detail.Arg2-as-detail* sense, so all *specification* relations connected to a list had to be checked manually to decide which of them can be interpreted also as explicit *Expansion.Level-of-detail.Arg2-as-detail*. From 82 *specification* relations connected with a list relation, 16 cases could be transformed to the corresponding Penn style relation.

3.7. Technical notes

From all technical parts of the transformation process, the extraction of arguments of the relations from their deep-syntactic tree representations to plain text proved to be the most challenging one. The numerous issues can be split in two categories: (i) annotation inconsistencies in various parts of the data (on the deep-syntactic layer, on the surface-syntactic layer, in the discourse annotation), and (ii) a complex nature of the deep-syntactic layer of annotation (reconstructed nodes/parts of the trees that take part in discourse relations, necessity to combine information from several annotation layers). Although we took great care in tuning the plain text generation of the arguments, we could not check and fix errors in all 21 thousand of discourse relations.

To demonstrate the kind of phenomena involved in discourse relations with elided (and reconstructed) nodes, consider Examples 13 and 14.

(13) ... *nechtěli* [*povolit*] nebo **nemohli odklad platby povolit** (PDiT, cmpr9410_002)
[... *would not* [*allow*] or **could not allow payment deferral**]

(14) *Celní unie bude existovat na papíře ještě dalších dvanáct měsíců* (a **třeba [bude existovat] i déle**) ... (PDiT, cmpr9410_001)
[*The customs union will exist on paper for another twelve months* (and **maybe [will exist] even longer**) ...]

In both cases, a discourse relation holds technically between two tectogrammatical nodes representing the same content verb, one of them being elided in the surface form of the sentence: *povolit* [*to allow*] in the first example and *existovat* [*to exist*] in the second example. In the first case, the actual discourse relation holds rather between the auxiliary verbs *nechtěli* [*would not*] and *nemohli* [*could not*], and although auxiliary nodes are not directly present at the tectogrammatical layer, they need to be represented in the plain text versions of the arguments. On the contrary, in the second

[17] From the list of implicit connectives – i.e. connectives filled in by annotators when annotating implicit relations – it seems that the Prague type list would be labeled as *Expansion.Conjunction*, because expressions *first*, *second*, *third* are listed there as connectives of implicit *Expansion.Conjunction* relations (PDTB 3.0 manual, Webber et al., 2019). However, expressions *first*, *second*, *third* are not listed in the list of explicit connectives, so this interpretation is just a guess.

case, the auxiliary node *bude* [*will*] needs to be present only in the first argument and omitted from the second one.[18]

Further, in the Prague style of discourse annotation, supplementary text parts were not annotated separately from the argument delimitation. Although the minimality principle was followed, in cases where the surrounding sentences played a distinct role in the discourse relation, they were marked as a part of the argument. In such cases, the additional sentences are transformed to the Penn style as supplementary texts.

Definitions of all data fields in the column format used for the transformed PDiT data are given in Table 3 in Appendix. Most of them come from the PDTB 3.0 data format; we have added a few fields to keep the original Prague discourse type and to provide plain text versions of information only captured in the form of spans in other fields.

4. Results and Discussion

Table 2 represents an overview of the result of the Prague discourse types to Penn senses transformation in the Prague Discourse Treebank data. The table shows a comparison of distributions of senses in (transformed) PDiT 3.0 and the PDTB 3.0 (in the latter taking into account explicit discourse relations only[19]). The two corpora are close to each other in size (both approx. 50 thousand sentences), genres (journalistic texts), in total numbers of explicit discourse relations (21 thousand vs. 25 thousand) and, as can be observed in the table, also in distributions of explicit discourse relations senses.

Although the sense frequencies in the two corpora are close in most of the cases, for several senses there are noticeable differences – they are highlighted in the table with grey background. Some of them may have roots in differences in the theoretical backgrounds of the two annotation styles, some others may simply reflect language or corpora differences. This constitutes a research question which inspired the following analysis. Let us elaborate below on the individual cases of noticeable differences in sense frequencies; for each sense, we state in parentheses the numbers of occurrences in the PDiT 3.0 transformed data and in the PDTB 3.0 data (but considering the slightly different total numbers of explicit relations in the two corpora, please take into account also the relative frequencies in the table).

[18] ...although it is referenced (via a link to the surface-syntactic layer) from both nodes representing the content verb *existovat* [*to exist*]. This can happen even in discourse relations between two non-elided content verbs, e.g. *Trámy byly urychleně rozebrány ...* a **[byly] odvezeny do dílen ...** (PDiT, ln94210_95) [*The beams were quickly disassembled* and **[they were] taken to the workshops ...**].

[19] i.e., in the PDTB terminology, relations marked as Explicit, AltLex and AltLexC

sense	PDiT	%	%	PDTB
Comparison.Concession.Arg1-as-denier	568	2.6%	2.9%	742
Comparison.Concession.Arg2-as-denier	3 551	16.4%	15.7%	4 057
Comparison.Concession+SA.Arg2-as-denier+SA	4	0.0%	0.1%	17
Comparison.Contrast	780	3.6%	4.5%	1 155
Comparison.Similarity	47	0.2%	0.7%	169
Contingency.Cause.Reason	1 750	8.1%	6.6%	1 712
Contingency.Cause.Result	1 299	6.0%	4.5%	1 160
Contingency.Cause+Belief.Reason+Belief	123	0.6%	0.1%	34
Contingency.Cause+Belief.Result+Belief	7	0.0%	0.0%	7
Contingency.Cause+SA.Reason+SA	2	0.0%	0.0%	1
Contingency.Cause+SA.Result+SA	4	0.0%	0.0%	1
Contingency.Condition.Arg1-as-cond	48	0.2%	0.1%	27
Contingency.Condition.Arg2-as-cond	1 237	5.7%	5.6%	1 445
Contingency.Condition+SA	102	0.5%	0.3%	73
Contingency.Negative-cause.NegResult	8	0.0%	0.0%	4
Contingency.Negative-condition.Arg1-as-negCond	2	0.0%	0.1%	16
Contingency.Negative-condition.Arg2-as-negCond	48	0.2%	0.4%	110
Contingency.Purpose.Arg1-as-goal	6	0.0%	0.5%	117
Contingency.Purpose.Arg2-as-goal	415	1.9%	1.2%	299
Expansion.Conjunction	8 161	37.8%	34.4%	8 907
Expansion.Disjunction	367	1.7%	1.2%	304
Expansion.Equivalence	127	0.6%	0.1%	37
Expansion.Exception.Arg1-as-excpt	6	0.0%	0.1%	15
Expansion.Exception.Arg2-as-excpt	195	0.9%	0.1%	24
Expansion.InstantiationArg1-as-instance	2	0.0%	0.0%	3
Expansion.InstantiationArg2-as-instance	206	1.0%	1.4%	375
Expansion.Level-of-detail.Arg1-as-detail	136	0.6%	0.2%	51
Expansion.Level-of-detail.Arg2-as-detail	646	3.0%	1.0%	262
Expansion.Manner.Arg1-as-manner	-	-	0.0%	3
Expansion.Manner.Arg2-as-manner	-	-	1.1%	280
Expansion.Substitution.Arg1-as-subst	61	0.3%	0.4%	111
Expansion.Substitution.Arg2-as-subst	391	1.8%	0.5%	137
Temporal.Asynchronous.Precedence	686	3.2%	4.1%	1 071
Temporal.Asynchronous.Succession	341	1.6%	4.5%	1 171
Temporal.Synchronous	262	1.2%	7.7%	1 981
total	21 588	100%	100%	25 878

Table 2. Comparison of distributions of senses in PDiT 3.0 and the PDTB 3.0. Please note that in the names of the senses, 'SpeechAct' was shortened to 'SA' to fit the page. Substantially different frequencies are highlighted with grey background.

Comparison.Similarity (47 in PDiT vs. 169 in the PDTB)

This difference results from different theoretical decisions. In the Prague style, all dependent clauses expressing manner were left out of the annotation, because they were considered not to be a separate abstract object and therefore did not form a discourse argument. Manner can be expressed also by comparison – and similarity is one type of comparison. Thus all cases of *Comparison.Similarity* in transformed PDiT come from discourse type of *conjunction* and do not appear in constructions with a dependent clause expressing manner by means of comparison.

Contingency.Cause+Belief.Reason+Belief (123 vs. 34)

Difference in the frequencies of this relation lies in our opinion in the fact that Czech has a special connective signalling this relation, connective *totiž* [*you see, actually*], which, besides other functions, can signal an argument for a claim. All examples of *Contingency.Cause+Belief* in the PDTB 3.0 manual (Webber et al., 2019) use details (not a reason) as evidence of justification for the presented claim and a majority of them are implicit (without a connective); in Czech, connective *totiž* is used in such contexts. Relations with this connective form 50 percent of all instances of *Contingency.Cause+Belief.Reason+Belief* in PDiT.

Contingency.Purpose.Arg1-as-goal (6 vs. 117)

Except for two cases, all instances of this relation in the PDTB 3.0 have connective *by* which can be in Czech expressed either by a dependent clause with connective *tím, že* [lit. *by that that*] or by a noun in the instrumental case (i.e. without any conjunction or preposition, without any connective) – none of these options is considered to be discourse relevant in the Prague style. Besides, these relations in the PDTB 3.0 hold mostly between arguments without finite verbs – as shown by Example 15. So this difference reflects both theoretical and language differences.

(15) *to correct this problem* by **providing a reliable flow of lendable funds** (PDTB, wsj_1131)

Expansion.Equivalence (127 vs. 37)

We could not find a satisfactory explanation for the different frequencies of this relation. It may be given by the polysemous nature of connective *tedy*, which corresponds to English *so*, *therefore*, but also to connective *in other words* and in some contexts more interpretations are possible. *Expansion.Equivalence* relations with connective *tedy* form 40 percent of all instances of this relation in PDiT.

Expansion.Exception.Arg2-as-excpt (195 vs. 24)

As described in detail in section 3.4, the PDiT relation of *restrictive opposition* corresponds partially to *Expansion.Exception* and at the same time includes also cases which would be interpreted as *Comparison.Contrast* in the PDTB 3.0 taxonomy. Manual analysis of the *restrictive opposition* relation in PDiT covered only the most frequent constructions and connectives, not all instances of the relation.

Expansion.Level-of-detail.Arg2-as-detail (646 vs. 262)

This difference stems from a theoretical decision to consider a colon and a dash to be discourse connectives in the Prague style – relations *Expansion.Level-of-detail.Arg2-as-detail* with these connectives form 60 percent of all instances of this relation in PDiT.

Expansion.Manner (0 vs. 283)

As already mentioned above, in the Prague style, clauses expressing manner were not considered to be separate abstract objects, so they were treated as a syntactic, not a discourse phenomenon.

Expansion.Substitution.Arg2-as-subst (391 vs. 137)

The higher frequency of this relation in PDiT is in our opinion given by the nature of the underlying PDT data – namely by the fact that elided verbs are reconstructed in the dependency trees of the deep-syntactic (tectogrammatical) layer of the corpus, thus allowing to annotate discourse relations with two verbal arguments in constructions such as *it is not A but B* (in Czech typically with an elided verb in B). For example, in the second part of the context in Example 16, there is a node for elided verb *poskytnout* [*to provide*]. Reconstructed nodes for elided verbs take part in annotation of 40 percent of all relations *Expansion.Substitution.Arg2-as-subst* in PDiT.

(16) *Tyto prostředky neposkytne místním spotřebitelům*, ale **[poskytne je] japonským zemědělcům.** (PDiT, ln94208_147)

[*It will not provide these funds to local consumers*, but **[it will provide them] to Japanese farmers.**]

Temporal.Synchronous (262 vs. 1981)

Upon close examination, we attribute this difference to a large extent to theoretical differences in the two annotation styles. Frequencies of translation counterparts of the most common connectives with this sense differ substantially. For example, whereas the most frequent Czech connective for this sense *když* has 783 occurrences in PDiT and only 100 of them are annotated as *Temporal.Synchronous*, its English counterpart *when* has 1 076 occurrences in the PDTB and half of them are assigned the *Temporal.Synchronous* sense (Webber et al., 2019). Besides, approx. 650 of the PDTB 3.0 *Temporal.Synchronous* relations have been labelled also by a second sense (*Comparison.Contrast*, *Contingency.Cause.Reason* etc.). As second senses are not annotated in the Prague style, sometimes other discourse types than temporal took precedence in the PDiT annotation if they were present in the given context. In contexts such as in Example 17, the Prague style would annotate just the *reason–result* relation, whereas the Penn style annotates *Temporal.Synchronous* as the first sense and *Contingency.Cause.Reason* as the second sense.

(17) *The company acquired the debt* when **it paid $155 million to purchase Wilson last year** (PDTB, wsj_0510)

5. Conclusion

The Prague Discourse Treebank data transformed to the Penn style of discourse annotation was published in December of 2022 in LINDAT/CLARIAH-CZ repository under the Creative Commons licence[20] as the Prague Discourse Treebank 3.0 (PDiT 3.0; Synková et al., 2022). The data was published in two formats: (i) the original Prague format of discourse annotation on top of tectogrammatical trees,[21] and (ii) the Penn column format of discourse annotation accompanied by the original plain texts. The discourse research community thus gets to its disposal another large-scale corpus manually annotated with discourse relations in the PDTB 3.0 style.

Understanding of the differences between the Prague and Penn semantic types taxonomies and of limits of the automatic transformation of the Prague discourse types to the Penn senses, based on a detailed study of both respective corpora, their annotation manuals and on a comparison of distributions of discourse relation senses in the two corpora, belong to the main theoretical results of the presented research. Frequencies of senses in the transformed PDiT data and in the PDTB 3.0 data are interestingly very similar. We have discussed the cases of senses where these frequencies considerably differed.

Differences in the taxonomies may in some cases reflect differences in the languages. For example, English has a particular connective – *unless* – for the relation of *Contingency.Negative-condition*, while Czech does not have its direct counterpart. During the conversion of the PDiT discourse annotation to the Penn style, we encountered a need to take a deeper look at how sentences corresponding to the English usage of

[20] `http://hdl.handle.net/11234/1-4875`

[21] For licensing reasons, the PDiT 3.0 distribution does not actually contain the tectogrammatical trees (and the lower layers of annotation); instead, the underlying data needs to be downloaded separately from the LINDAT/CLARIAH-CZ repository (the PDT part of the PDT-C 1.0, `http://hdl.handle.net/11234/1-3185`) and the discourse annotation can be added to the data by a script provided by the PDiT 3.0 distribution.

unless are constructed in Czech. In this way, we found certain tendencies combining the use of a particular connective, sentence negation and the position of discourse arguments.

The theoretical results are reflected also in technical procedures developed during the presented research for transforming the Prague style of discourse annotation to the Penn style. These procedures can be used in future for any data annotated in the Prague style of discourse annotation. They consist of two separate parts: (i) transformation of discourse arguments and connectives from their representation in tectogrammatical trees to plain text, and (ii) transformation of Prague discourse types to Penn senses.

Thousands of discourse relations in the PDiT data were examined during the research, resulting in many rules embedded in the transformation procedures. These rules were used to transform discourse types of 54 percent of all PDiT discourse relations (12 thousand out of over 21 thousand). 42 percent (over 9 thousand) of the PDiT relations carry a discouse type that transforms to a single Penn sense; their discourse types were also transformed automatically. In the end, discourse types of only 1.8 percent of all discourse relations in the PDiT data (388 relations) had to be disambiguated manually in order to be transformed to the correct sense.

The project covering this research will continue for two more years, having as its ultimate goal to have the whole Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)[22] annotated with discourse relations and published in both the Prague and Penn styles of discourse relations annotation.

Acknowledgement

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 (LM2018101).

Bibliography

Bojar, Ondřej, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. Evald Reference-less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 541–545, 2018. doi: 10.18653/v1/W18-6432.

Danlos, Laurence, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. Vers le FDTB: French Discourse Tree Bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 471–478, 2012.

[22] The Prague Dependency Treebank - Consolidated 1.0 consists of four subcorpora: (i) the PDT, (ii) the Czech part of the Prague Czech-English Dependency Treebank, (iii) the Prague Dependency Treebank of Spoken Czech, and (iv) Faust. Altogether, PDT-C 1.0 includes approx. 175 thousand sentences.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of NAACL: Human Language Technologies*, pages 4171–4186, 2019.

Hajič, Jan, 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. Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0), 2020.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. *Prague Czech-English Dependency Treebank 2.0*. Data/Software, Linguistic Data Consortium, 2012. University of Pennsylvania, Philadelphia. LDC2012T08.

Kiddon, Chloé, Luke Zettlemoyer, and Yejin Choi. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, 2016. doi: 10.18653/v1/D16-1032.

Lee, Alan, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Annotating Discourse Relations with the PDTB Annotator. In *Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: System Demonstrations*, pages 121–125, 2016.

Mann, William C. and Sandra A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281, 1988. doi: 10.1515/text.1.1988.8.3.243.

Meyer, Thomas and Bonnie Webber. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, 2013.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Technical report, ÚFAL MFF UK, Prague, 2005.

Mírovský, Jiří and Lucie Poláková. Sense Prediction for Explicit Discourse Relations with BERT. In Yang, Xin-She, Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Sixth International Congress on Information and Communication Technology (ICICT)*, volume 216 of *Lecture Notes in Networks and Systems*, pages 835–842, Singapore, 2021. International Congress and Excellence Awards, Springer. ISBN 978-981-16-1781-2.

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. Does Tectogrammatics Help the Annotation of Discourse? In *Proceedings of COLING 2012: Posters*, pages 853–862, 2012. URL `https://www.aclweb.org/anthology/C12-2083.pdf`.

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. Discourse Relations in the Prague Dependency Treebank 3.0. In Tounsi, Lamia and Rafal Rak, editors, *The 25th International Conference on Computational Linguistics (Coling 2014), Proceedings of the Conference System Demonstrations*, pages 34–38, Dublin, Ireland, 2014. Dublin City University (DCU), Dublin City University (DCU).

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

Oza, Umangi, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. The Hindi Discourse Relation Bank. In *Proceedings of the third Linguistic Annotation Workshop*, pages 158–161, 2009. doi: 10.3115/1698381.1698410.

Pajas, Petr and Jan Štěpánek. Recent Advances in a Feature-rich Framework for Treebank Annotation. In Scott, Donia and Hans Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester, 2008. The Coling 2008 Organizing Committee. doi: 10.3115/1599081.1599166. URL `https://www.aclweb.org/anthology/C08-1085.pdf`.

Poláková, Lucie and Pavlína Synková. Pragmatické aspekty v popisu textové koherence. *Naše řeč*, 104(4):225–242, 2021. ISSN 0027-8203.

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

Poláková, Lucie, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, 2013. Asian Federation of Natural Language Processing.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, 2008. European Language Resources Association. URL `http://lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf`.

Prasad, Rashmi, Bonnie Webber, Alan Lee, and Aravind Joshi. *Penn Discourse Treebank Version 3.0*. Data/Software, Linguistic Data Consortium, 2019. URL `https://catalog.ldc.upenn.edu/LDC2019T05`. University of Pennsylvania, Philadelphia. LDC2019T05.

Rysová, Kateřina, Magdaléna Rysová, and Jiří Mírovskỳ. Automatic Evaluation of Surface Coherence in L2 Texts in Czech. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016)*, pages 214–228, 2016.

Rysová, Magdaléna and Kateřina Rysová. The Centre and Periphery of Discourse Connectives. In *Proceedings of Pacific Asia Conference on Language, Information and Computing*, pages 452–459, Bangkok, 2014. Department of Linguistics, Faculty of Arts, Chulalongkorn University. URL `https://www.aclweb.org/anthology/Y14-1052.pdf`.

Rysová, Magdaléna and Kateřina Rysová. Primary and secondary discourse connectives: Constraints and preferences. *Journal of Pragmatics*, 130:16–32, 2018. ISSN 0378-2166. doi: 10.1016/j.pragma.2018.03.013.

Rysová, Magdaléna, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, and Šárka Zikánová. Prague Discourse Treebank 2.0. Data/Software. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2016. URL `http://hdl.handle.net/11234/1-1905`.

Shi, Wei and Vera Demberg. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5794–5800, 2019. doi: 10.18653/v1/D19-1586.

Synková, Pavlína, Magdaléna Rysová, Jiří Mírovský, Lucie Poláková, Veronika Sheller, Jana Zdeňková, Šárka Zikánová, and Eva Hajičová. Prague Discourse Treebank 3.0, 2022.

Taboada, Maite and William C Mann. Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse studies*, 8(3):423–459, 2006. doi: 10.1177/1461445606061881.

Turney, Peter D and Michael L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003. doi: 10.1145/944012.944013.

Webber, Bonnie, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545–587, 2003. doi: 10.1162/089120103322753347.

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

Xiong, Hao, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling Coherence for Discourse Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345, 2019. doi: 10.1609/aaai.v33i01.33017338.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 1–16, 2015. doi: 10.18653/v1/K15-2001.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proc. of the CoNLL-16 shared task*, pages 1–19, 2016. doi: 10.18653/v1/K16-2001.

Zeyrek, Deniz and Murathan Kurfalı. TDB 1.1: Extensions on Turkish Discourse Bank. *LAW XI 2017*, page 76, 2017. doi: 10.18653/v1/W17-0809.

Zhang, Renxian. Sentence Ordering Driven by Local and Global Coherence for Summary Generation. In *Proceedings of the ACL 2011 Student Session*, pages 6–11, 2011.

Zhou, Yuping and Nianwen Xue. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the ACL: Long Papers-Volume 1*, pages 69–77, 2012.

Zikánová, Šárka, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia, 2015. ISBN 978-80-904571-8-8.

Zikánová, Šárka, Pavlína Synková, and Jiří Mírovský. Enriched Discourse Annotation of PDiT Subset 1.0 (PDiT-EDA 1.0), 2018.

Appendix

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

Besides the original PDiT format of the data, the transformed discourse annotation is also provided in the PDTB 3.0 column text format where each 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. For compatibility reasons, we have kept all field definitions from the PDTB 3.0 (although not all of them are actually used in the transformed PDiT data[23]) and for additional information, we have added new fields. Table 3 gives field definitions of the format used for the PDiT transformed data. The first part of the table, fields 0–33, corresponds to the original PDTB 3.0 fields; it is taken from the PDTB 3.0 annotation manual (Webber et al., 2019) and the definitions are adjusted to better fit our data. The second part, fields 34–43, gives definitions of additional fields introduced in the PDiT 3.0 transformed data.

Address for correspondence:
Jiří Mírovský
mirovsky@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic

[23] Neither in the PDTB 3.0 are all of them used, as the PDTB 3.0 keeps backward format compatibility with its previous version.