Cost-Effective Discourse Annotation in the Prague Czech–English Dependency Treebank

Jiří Mírovský, Pavlína Synková, Lucie Poláková and Marie Paclíková
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic
{mirovsky, synkova, polakova}@ufal.mff.cuni.cz, marie.paclikova@gmail.com

Abstract

We present a cost-effective method for obtaining a high-quality annotation of explicit discourse relations in the Czech part of the Prague Czech–English Dependency Treebank, a corpus of almost 50 thousand sentences coming from the Czech translation of the Wall Street Journal part of the Penn Treebank. We use three different sources of information and combine them to obtain the discourse annotation: (i) annotation projection from the Penn Discourse Treebank 3.0, (ii) manual tectogrammatical (deep syntax) representation of sentences of the corpus, and (iii) the Lexicon of Czech Discourse Connectives CzeDLex. After solving as many discrepancies as possible automatically, the final discourse annotation is achieved by manual inspection of the remaining problematic cases. The discourse annotation of the corpus will be available both in the Prague format (on top of tectogrammatical trees) with the Prague taxonomy of discourse types, and in the Penn format (on plain texts) with the Penn Discourse Treebank 3.0 sense taxonomy.

Keywords: discourse relations, annotation projection, cost-effective annotation, Prague Czech–English Dependency Treebank

1. Introduction

Creating a high-quality discourse-annotated corpus is a resource-demanding task and, although many discourse-annotated corpora have been developed in the last years for more than ten different languages, only a few of them are of significant size. In the area of shallow discourse analysis, the Penn Discourse Treebank (PDTB) with approx. 26 thousand explicit discourse relations (Prasad et al., 2019) and the Prague Discourse Treebank (PDiT) with approx. 22 thousand (explicit) discourse relations (Synková et al., 2022) belong to the largest ones. And yet, not only the quality but also the size of a corpus is important for its usability both in theoretical research and in NLP applications (e.g., Kocián et al., 2022; Zeman, 2004 as examples of the omnipresent data-sparsity problem in manual annotations both in post-BERT and pre-BERT eras), and a whole field of research is dedicated to determining the amount of manually annotated data needed for training to achieve a certain model performance for a given task (Chang et al., 2023; Lauer, 1995).

It is therefore beneficial to devise methods of rapid and cost-effective linguistic annotation, employing all available language resources. Such an opportunity emerged for discourse relations in Czech thanks to the existence of several key elements:

1. a parallel English–Czech corpus,
2. a manual discourse annotation on the English part of the corpus,
3. a discourse parser for Czech,
4. a manual deep-syntax annotation of the Czech part of the corpus, and
5. a richly annotated lexicon of Czech discourse connectives.

We used a combination of these resources to get a high-quality discourse annotation of the Czech part of the Prague Czech–English Dependency Treebank, automatizing most of the work and minimizing the necessary human intervention.

At the heart of our method, there is an annotation projection of discourse relations from English to Czech. Annotation projection is a well known method of inducing annotation of a phenomenon in one language providing that such an annotation already exists on parallel texts in another language. Annotation projection has been widely used also in the area of discourse relations, for example Versley (2010) uses annotation projection from automatically discourse-parsed English to German in an English–German parallel corpus. Laali and Kosseim (2017) used annotation projection to get a discourse annotated corpus for French from Europarl, proposing a novel approach to filter out unsupported annotations. In Mírovský et al. (2021), we used annotation projection from English to Czech (on the same data as we did in the present paper) to get corpus examples of discourse connective usages to enrich a lexicon of Czech discourse connectives.

The paper is organized as follows. In Section 2, we introduce all relevant language resources. Section 3 describes the individual steps leading to the final discourse annotation. Section 4 mentions the benefits that a suitable annotation tool can bring to the annotation process, and Sections 5 and 6 evaluate and summarize the results of the work.

2. Relevant Resources

PCEDT 2.0: The Prague Czech–English Dependency Treebank 2.0 (PCEDT) (Hajič et al., 2012) is a corpus of English–Czech parallel texts (50 thousand sentences at each side) and their analyses on several layers of language description. The original – English – texts come from the Wall Street Journal section of the Penn Treebank (PTB) (Marcus et al., 1995) and have been human-translated to Czech, keeping 1:1 sentence correspondence (Hajič et al., 2012). The manual analysis of the texts on both language sides goes up to the deep syntax (tectogrammatical) layer of language description, where the sentence is represented by a dependency tree with content words as nodes (see Sgall et al., 1986 for the theoretical background of the Functional Generative Description). From the point of view of the present paper, it is important that the tectogrammatical layer largely allows for (i) recognition of discourse arguments (the trees also contain nodes for elided verbs), (ii) recognition of connectives (they have defined positions in the tree according to their types), and (iii) interpretation of the type of intra-sentential discourse relations using semantic labels called functors.

The corpus provides automatic alignment on all its layers, i.e., for example, for tokens on the word layer, and for tree nodes on the tectogrammatical layer.[1]

PDTB 3.0: The Penn Discourse Treebank 3.0 (Prasad et al., 2019) is the most recent iteration of manual discourse annotation of (most of) the texts of the Wall Street Journal part of the Penn Treebank, i.e. the texts that also form the English part of the PCEDT. For our purposes, it is important that the discourse annotation includes approx. 26 thousand explicit relations (incl. so called alternative lexicalizations), i.e. discourse relations marked in the text by a connective. Discourse annotation in the PDTB is done on plain texts and in the past was several times mapped onto the dependency trees of the English part of the PCEDT (e.g. Mírovský et al., 2021).

CzeDLex 1.0: CzeDLex is both a human- and machine-readable online dictionary of Czech discourse connectives (Mírovský et al., 2021).[2] It includes over 200 basic entries (connectives such as *ale* [*but*], *proto* [*therefore*]) and their common complex forms (e.g. *a proto* [*and therefore*]) and modifications (e.g. *možná proto* [*maybe therefore*]). Each entry is complemented with additional information: corpus frequencies, usages in various discourse types with examples, argument semantics with respect to the position of the connective etc. (Mírovský et al., 2017). CzeDLex was devised (and manually post-edited) from several discourse-annotated resources, mainly from the Prague Discourse Treebank 2.0 (Rysová et al., 2016) but also from an annotation projection from the PDTB (via the English–Czech PCEDT), see Mírovský et al. (2021).

3. Method

The whole method consists of the following seven steps, which we elaborate in the subsequent subsections.

1. mapping plain text PDTB 3.0 discourse relations to the tectogrammatical trees in the English part of the PCEDT (PCEDT-en)
2. projecting the PDTB relations to the Czech part of the PCEDT (PCEDT-cs)
3. discourse-parsing the PCEDT-cs
4. merging the parsing with the projection
5. solving discrepancies regarding the existence of relations
6. solving discrepancies in Prague vs. Penn discourse types/senses[3]
7. solving ambiguous transformation to Penn senses

3.1. PDTB 3.0 → PCEDT-en

The first step was mapping the PDTB 3.0 annotation (which is in a form of stand-off indexing to plain text) on the tectogrammatical trees of the English part of the PCEDT. Such a transformation has been done before and described in literature; we proceeded in the same way as we did in Mírovský et al. (2021).

3.2. PCEDT-en → PCEDT-cs

In the subsequent step, the discourse relations were projected from the English tectogrammatical trees to their Czech counterpart, using automatic alignments on tokens and on tectogrammat-

[1] The alignment on individual layers in the PCEDT is based on GIZA++ tool, see, e.g., Mareček et al., 2008.

[2] https://ufal.mff.cuni.cz/czedlex1.0

[3] *Discourse type* refers to a semantico-pragmatic type of a discourse relation in the Prague taxonomy, while *sense* refers to the semantico-pragmatic type of a discourse relation in the PDTB 3.0 (Penn) taxonomy.

ical nodes, as provided by the PCEDT.[4] Again, we proceeded in a way similar to the one we had already described in Mírovský et al. (2021). Given the imperfection of the automatic alignment and because of expression differences arising from the translation of the text, the result of the projection was a very rough and erroneous approximation of discourse relations in the Czech text.[5]

3.3. Discourse Parsing

As a second source of discourse relations annotation in the Czech part of the PCEDT, we used an updated version of our own Czech shallow discourse parser. The parser combines information from the manual annotation of the deep-syntax (tectogrammatical) layer of the texts with information from the Lexicon of Czech Discourse Connectives CzeDLex.

The annotation on the tectogrammatical layer allows the parser (to a certain extent) to recognize and classify intra-sentential relations from the structure of the dependency trees and from the labels ("functors") attached to the edges of the trees. For example (see Figure 1 and Example 1), if a tectogrammatical edge has the COND functor ("condition") and it connects two nodes representing finite verbs (two roots of tensed clauses: in our case, *přistoupit* [*to approach*] and *mít* [*to be*]),[6] it is used by the parser to annotate a discourse relation between the dependent and main clauses represented by the two verb nodes and their subtrees. The connective *jestliže* [*if*] is searched for among auxiliary nodes[7] linked from the dependent verb node *přistoupit* [*to approach*] and crosschecked with CzeDLex. The discourse type *condition* is assigned based on probabilities of discourse types usually derived from the COND functor (this information comes from PDiT) and on probabilities of discourse types associated with connective *jestliže* [*if*] (this information comes from CzeDLex). Although the tectogrammatical annotation does not help in classifying the discourse type of inter-sentential relations, most connectives for inter-sentential relations can be identified using information from the tectogrammatical trees (their position in the tree and their functor)[8] and from CzeDLex. For classifying the discourse type of inter-sentential relations, the parser relies entirely on CzeDLex.

3.4. Merging Projection with Parsing

As a result of the previous steps, two different kinds of discourse relations were captured in the Czech PCEDT data: the ones projected from the PDTB and the parsed relations. Both types are represented in the trees as arrows (of different color and style) between two nodes that represent the arguments of the relation (most often, an argument corresponds to the subtree of the representative node). For simplicity, we refer to the PDTB-projected relations as *PDTB relations* or *PDTB arrows* and to the parsed relations as *AUTO relations* or *AUTO arrows*.

In this step, we merged PDTB and AUTO arrows that were likely to express the same relation. In the ideal case, both start and target nodes of the arrows were identical, the discourse type of the AUTO arrow was compatible with the respective sense of the PDTB arrow[9] and the connectives attached to the arrows were adequate equivalents; see, e.g., Figure 1 representing Example 1.[10]

(1) Jestliže **k tomu přistoupí velkoryse a nesobecky**, *budou mít čistý zisk všichni*. (PCEDT, wsj0043)

[If **they approach it with a benevolent, altruistic attitude**, *there will be a net gain for everyone*.]

Such arrows were automatically merged in the following way: the PDTB arrow was deleted, its sense was copied to a special attribute at the AUTO arrow and the AUTO arrow was marked as

[4] Although all types of relations were projected (incl. Implicit ones), only Explicit, AltLex and AltLexC arrows were used in the subsequent merge.

[5] In the mentioned paper, we used the resulting discourse relations to extract discourse connectives along with semantic senses of their usages to enrich the Lexicon of Czech Discourse Connectives CzeDLex; only those discourse relations that represented a usage of a connective with a sense that had not been covered by CzeDLex previously were manually checked.

[6] The translation is not literate, yet the Czech word *mít* corresponds to the word *to be* in the English sentence.

[7] on the surface syntax (analytical) layer, not visible in the figure

[8] Most inter-sentential connectives carry one of the following functors: PREC (reference to preceding context), RHEM (rhematizer), TWHEN (temporal modification) or ATT (atomic expression of the speaker's attitude), sometimes also REG (expressing a circumstance), MEANS (expressing a means) and other temporal functors.

[9] Compatibility of discourse types and senses has been determined by studying annotation manuals of the two respective approaches (Prague and Penn) and their actual co-existence in the data of the Prague Discourse Treebank 3.0 (Synková et al., 2024, in print).

[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.

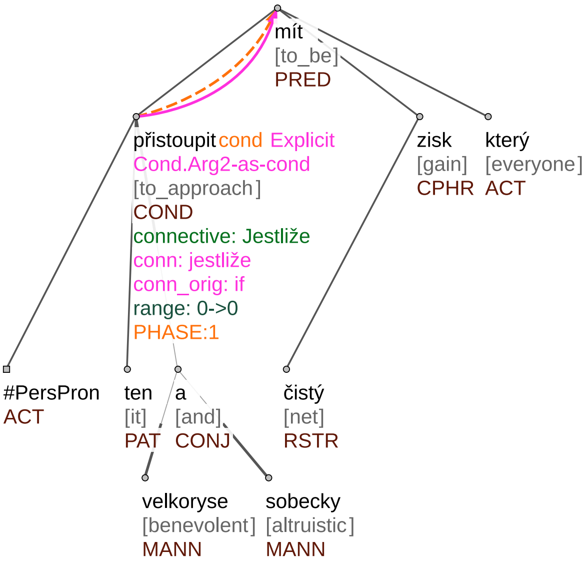

Figure 1: A perfect overlap from Example 1; the orange (dashed) arrow and the magenta (full-line) arrow represent the AUTO and PDTB relations, respectively. The automatically assigned discourse type (cond, *condition*) corresponds to the projected sense Condition.Arg2-as-condition, and the automatically detected connective (*Jestliže* [*If*]) is equal to the projected one (*jestliže*). The comment (PHASE:1) indicates that the AUTO arrow has not been confirmed yet. Please note that the translations of lemmas at the nodes are not part of the data and have been added here.

PHASE:2 along with a description of the type of the merge configuration (in the attribute *comment* of the arrow).

For less clear configurations, samples of 100 to 200 cases were examined and rules were devised for their full or at least partial automatic processing. Thus, for example, also intra-sentential arrows that agreed in both start and target nodes could be merged if at least one of the following conditions was met:

1. there was no other arrow starting or ending in the two nodes,

2. the arrows had a compatible discourse type vs. sense, or

3. the arrows had equal or at least overlapping connectives (e.g., *proto* [*therefore*] vs. *hlavně proto* [*mainly therefore*]).

Otherwise human inspection was needed to check which arrows actually belonged together.

A similar set of rules was devised for intra-sentential AUTO and PDTB arrows that differed either in their start or target nodes. Additional conditions for merging in these cases checked which of the nodes represented finite verbs (or, e.g., their coordination), as many errors were caused by errors in the English–Czech alignment.

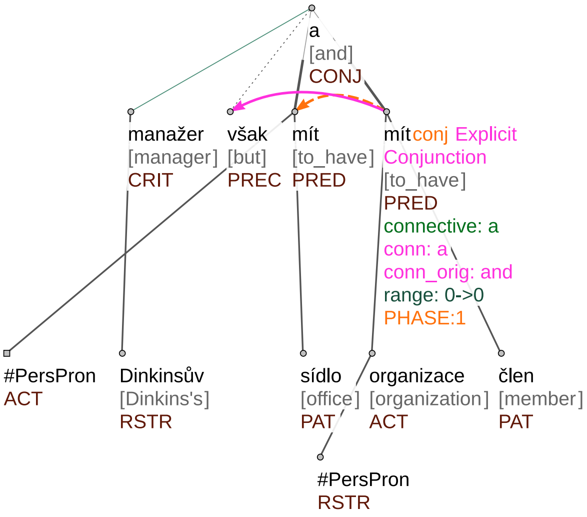

Figure 2: Different target nodes of the AUTO and PDTB arrows for the intra-sentential relation from Example 2.

If a less clear pair of arrows was identified as expressing the same relation (given the connectives, the relation types, presence of other arrows in the vicinity), the information from the PDTB arrow was still preserved to some extent – e.g., if the Penn sense and the Prague discourse type were compatible, the Penn sense was stored (and considered correct). If they were not compatible, it was stored as well but marked for manual inspection.

For example, the intra-sentential relations in Example 2 could be merged even if the target nodes of the arrows differed, see Figure 2 where an error in alignment caused a wrong projection of the target node of the PDTB arrow.

(2) Podle Dinkinsových manažerů však *měl sídlo* a **jeho organizace měla členy**. (PCEDT, wsj0041)

[But, say Mr. Dinkins's managers, *he did have an office* and **his organization did have members**.]

For intra-sentential relations, the identification of arguments (represented by start and target nodes) of AUTO arrows was usually reliable (except for sentences with reported speech, see below in Section 3.5), as the parser took advantage of the manual annotation of the tectogrammatical tree and (also manually annotated) morphological information of the nodes, while the PDTB arrows suffered from wrong alignment.

Similarly to the intra-sentential situation, various configurations of AUTO and PDTB arrows were studied for inter-sentential cases and sets of rules were devised to process parts of these cases auto-

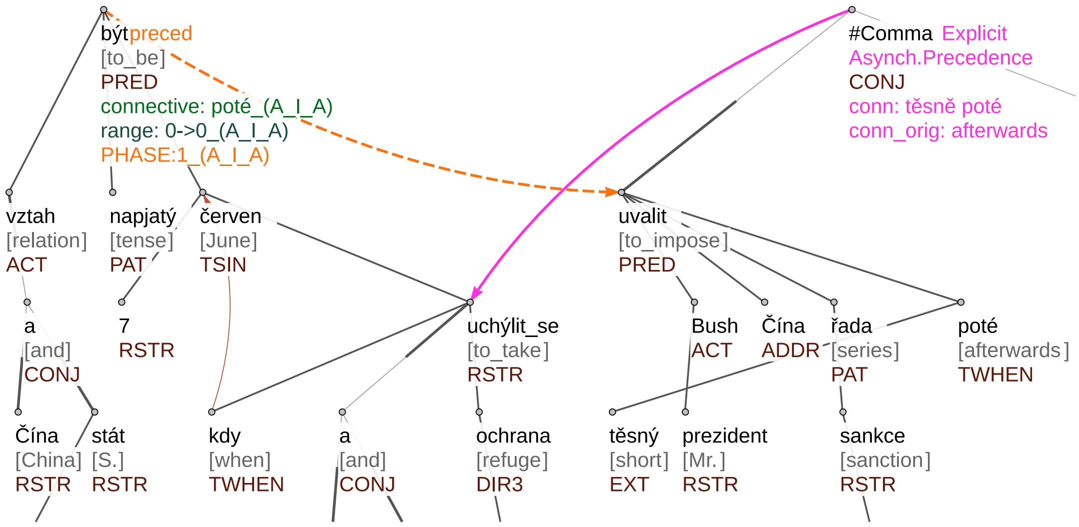

Figure 3: Relevant parts of the tectogrammatical trees of sentences from Example 3; the dashed AUTO arrow can be merged with the solid-line PDTB arrow, with *uchýlit_se* [*to_take*] and the *#Comma* as new start and target nodes, respectively.

matically. Here the parser could not rely as much on the structure of the tectogrammatical trees and the PDTB arrows proved more reliable for detecting less standard positions of the arguments.[11]

Example 3 and Figure 3 demonstrate a case where the inter-sentential PDTB and AUTO arrows differ both in the start and target nodes but they have compatible senses and connectives and there are no other inter-sentential arrows to mix them with. The arrows therefore could be merged automatically and, as the PDTB arrow connects a finite-verb node with a coordination of finite verbs (so no obvious alignment error had been detected), the merging procedure chose these nodes for the resulting arrow. The arguments and the connective marked in the text of Example 3 correspond to the result of the merge.

(3) Vztahy mezi Čínou a Spojenými státy jsou napjaté od 7. června, kdy *se čínský disident Fang Lizhi a jeho žena Li Shuxian uchýlili pod ochranu velvyslanectví Spojených států v Pekingu*. Těsně poté **uvalil prezident Bush na Čínu řadu sankcí, včetně přerušení rozhovorů na nejvyšší úrovni, což by mohlo být americkým Kongresem v nadcházejících týdnech kodifikováno v legislativě.** (PCEDT, wsj0093)

[Relations between China and the U.S. have been tense since June 7, when *Chinese dissident Fang Lizhi and his wife, Li Shuxian, took refuge in the U.S. Embassy in Beijing*. Shortly afterwards, **Mr. Bush imposed a series of anti-China sanctions, including suspension of most high-level talks, which could be codified in U.S. congressional legislation in the coming weeks.**]

The PDTB arrows were also the only available source for automatically setting correct arguments for distant inter-sentential relations or relations with arguments longer than one sentence, as the parser always creates inter-sentential relations between subsequent sentences.

Altogether in this part of the procedure, approx. 12 thousand pairs of intra-sentential AUTO and PDTB arrows and approx. 6 thousand pairs of inter-sentential AUTO and PDTB arrows could be merged automatically.

3.5. Discrepancies in Relation Identification

Even with loosened conditions for merging, many arrows (both PDTB and AUTO) remained in the data that could not be merged automatically without compromising the quality of the merge. Such cases were sorted according to the type of discrepancy and manually inspected (and solved) by experienced annotators. The discrepancies can be divided into two basic types.

First, there were cases where an AUTO arrow or a PDTB arrow stood alone, i.e., for intra-sentential relations, such an arrow was not accompanied in the same tree by another intra-sentential PDTB or AUTO arrow, respectively; or, for inter-sentential relations, there was no inter-sentential PDTB or AUTO arrow starting or ending in the same tree. These cases were mainly caused by:

(i) translation: consider Example 4, where English connective *and* has no equivalent in the

[11] As long as their start and target nodes represented finite verbs (or their coordination...).

Czech translation.[12]

(4) Puklinami na kraji silnice prorůstá plevel, mnoho domů a skladů je prázdných. (PCEDT, wsj1760)

[*Weeds push up through the cracks in the sidewalks* and **many houses and storefronts are empty**.]

(ii) differences in the linguistic characteristics of Czech and English: i.e. cases where Czech needs to use a clause where English allows the use of non-verbal expressions; some expressions function as connectives in English but have no connective function in Czech. A typical case is the expression *skutečně* [*indeed*], which in Czech only emphasizes a content and does not open two slots for arguments, while in English it often functions as a connective in inter-sentential relations, consider Example 5.[13]

(5) Mary Elizabeth Ariailová [...] se domnívala, že si kolegyně Yearginová chtěla udržet dobrou pozici, aby mohla získat novou práci, která nevyžadovala tak dobrý sluch. Yearginová se skutečně zajímala o případné zaměstnání v souvislosti se státním učitelským kadetním programem. (PCEDT, wsj0044)

[*Mary Elizabeth Ariail [...] believed Mrs. Yeargin wanted to keep her standing high so she could get a new job that wouldn't demand good hearing.* Indeed, **Mrs. Yeargin was interested in a possible job with the state teacher cadet program.**]

(iii) differences between the PDTB and Czech discourse theoretical frameworks: a PDTB argument can be represented by a gerund or an infinitive, while an argument in the Prague approach must include a finite verb; AltLexes (alternative lexicalizations of connectives) are considered much broader in the PDTB than secondary connectives in Czech; the PDTB also considers some relations to be a part of discourse, which the Prague system considers to be a part of syntax. The last case can be illustrated by Example 6, where the dependent clause was considered to be an argument of the condition relation in the PDTB, while in the Prague approach it is a part of the valency frame of the main verb, and as such not relevant for discourse annotation.

(6) Je zkrátka komické, když se snaží předstírat, že jsou stále nadřazenou rasou. (PCEDT, wsj0296)

[*It's just comic* when **they try to pretend they're still the master race**.]

(iv) atypical structures: PDTB arrows also appear alone in constructions where the parser did not recognize the relation because the syntactic structure was non-typical. This concerns clauses depending on an infinitive governed by a finite verb, or cases where one argument is a relative clause, as in Example 7, *z něhož však sešlo* [lit. *which was however called off*] in the Czech translation.

(7) *Vláda strávila většinu loňského roku snahou realizovat takový plán*, **z něhož** však **sešlo, když mateřský koncern Waertsilae na poslední chvíli vycouval**. (PCEDT, wsj0773)

[*The government spent most of last year attempting to carry out such a plan* but **was thwarted when the parent Waertsilae concern pulled out at the last minute**.]

The second type of discrepancies concerned cases where AUTO and PDTB arrows were both in the same tree or shared the tree where they started or ended, but the automatic procedure could not recognize which of the arrows (if any) were relevant in the given context. This situation was caused, besides the ubiquitous errors in the annotation projection,[14] also by (i) theoretical frame differences (mainly by the fact that the expressions as *and then* or *but also* are each considered two connectives for two relations in the PDTB, whereas in the Prague system they are considered complex connectives for a single relation), and (ii) by attribution in contexts with reported speech. Such a context is given in Example 8, where the AUTO arrow connects verbs in the main clauses, whereas the PDTB arrow connects verbs in the dependent clauses and represents the appropriate annotation of the case (as depicted in Figure 4).

(8) Představitelé uvedli, že *si nebyli jisti, jak budou tyto peníze mezi zahraniční jednotky rozděleny*, ale dodali, že **NEC Semiconductors U. K. Ltd. bude mít přednost**. (PCEDT, wsj0379)

[Officials said *they weren't sure how the money will be distributed among overseas units*, but added that **NEC Semiconductors U.K. Ltd. will receive priority**.]

[12] A comma by itself is not considered a connective.

[13] If *indeed* were just before *interested* here in English (or some more naturally sounding equivalent – the Czech word *skutečně* can be translated also as *really* here), we would – just as in Czech – not consider it to function as a connective (but we are not entirely sure what the authors of the PDTB would do here).

[14] caused by errors in the alignment

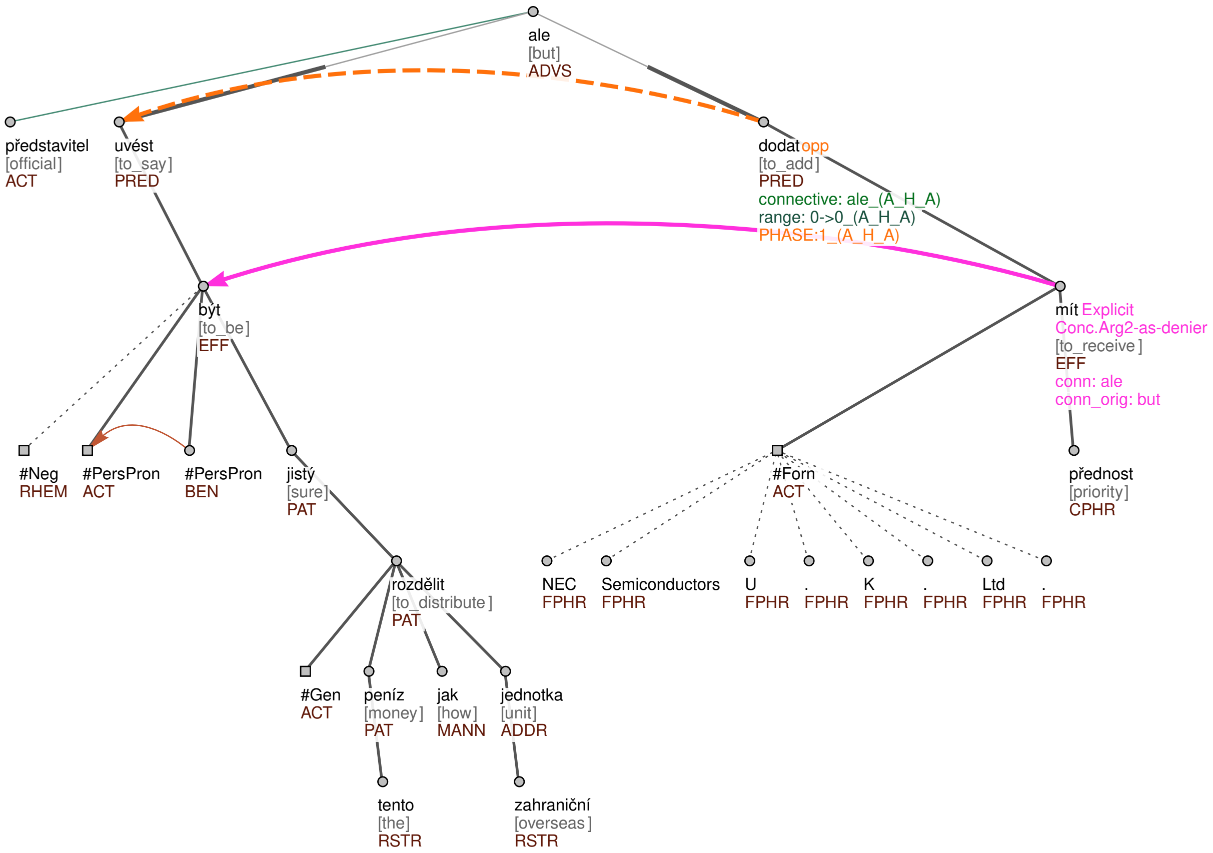

Figure 4: A relation between two attributed contents from Example 8; the dashed AUTO arrow is wrongly placed, the solid-line PDTB arrow connects the correct arguments.

In total, approx. 6 thousand cases were checked manually; the remaining stand-alone intra-sentential AUTO arrows (5 thousand cases) were considered correctly placed and without a need for inspection.

The manual checks resulted in the annotation of 3.5 thousand relations, of which 7 hundred could be taken completely from the projection (i.e., neither the arguments, the sense, nor the connective needed to be changed). The remaining relations were either taken directly from the AUTO arrows or they represented cases where the arrows from the projection had to be modified.

3.6. Discrepancies in Types/Senses

Merging a projected arrow with a parsed arrow sometimes resulted in a relation with a discrepancy between the Prague discourse type and the PDTB sense, i.e. the type and sense did not correspond to each other (for example, sense Contingency.Cause.Reason and type *conjunction*). The correspondence between the Prague discourse types and the Penn senses has been studied thoroughly in the past and the transformation process from the types to the senses is described in detail in Mírovský et al. (2023).

In general, the projected senses were helpful for relations with less common discourse types of the given connective. Consider, for example, the sentence in Example 9. The connective *avšak* [*however*] most often signals *opposition* (and the parser annotated it here), but in this particular context it signals Comparison.Contrast in the Penn taxonomy, i.e. *confrontation* in the Prague taxonomy.

(9) *Soukromé stavební výdaje klesly*, avšak **vládní stavební aktivity stouply.** (PCEDT, wsj0036)

[*Private construction spending was down*, but **government building activity was up.**]

On the other hand, automatically assigned types were more relevant especially when the translation affected the meaning. In Example 10, the connective *and* was translated as *a tak* [*and so*] and sense Expansion.Conjunction was thus not relevant for the Czech text.

(10) *Nebyla školená*, a tak **v jednom špatně provedeném zákroku způsobila smrt klientky.** (PCEDT, wsj0039)

[*She was untrained* and, **in one botched job killed a client.**]

Some discrepancies were also caused by differencies in exact definitions of the Prague discourse types and the Penn senses in the respective annotation systems. For example, a relation in Example 11 is considered *disjunctive alternative* in the Prague approach, while the Penn system annotates Contingency.Negative-condition.

(11) ...společnost Viacom Inc., jim dává ultimátum: buď *podepíší další dlouhodobý závazek k nákupu dalších epizod*, nebo **jim hrozí, že jim „Cosbyho" vezme konkurence**. (PCEDT, wsj0060)

[Viacom Inc., is giving an ultimatum: Either *sign new long-term commitments to buy future episodes* or **risk losing "Cosby" to a competitor**.]

Manual inspections of discrepancies between senses and discourse types, or inspection of ambiguous cases where no sense was available, were carried out in approx. 8 thousand positions in the data. The most frequent discrepancies that required manual checking arose from the Comparison class of relations and also from the fact that many relations annotated as Temporal.Synchronous in the PDTB 3.0 are not considered primarily temporal in the Prague approach.

3.7. Ambiguous Type → Sense

The last step requiring manual edits was the transformation of the Prague discourse types to the Penn senses, as the final corpus offers discourse annotation in both systems. Such a transformation was studied and described before on the data of the Prague Discourse Treebank (PDiT; the study Mírovský et al., 2023); according to the study, most of the discourse types translate to a sense unambiguously; they represented about 42% of relations occurring in the PDiT texts. 56% of relations could be transformed using rules based on linguistic features, and only about 2% had to be manually disambiguated.

In our present task of transforming the Prague discourse types to the Penn senses in the PCEDT-cs data, in most cases the disambiguation was already taken care of during the previous step (see Section 3.6), and the original PDTB sense from the projection helped with most of the problematic pragmatic discourse types and the relation of *explication*.

4. Tool

Cost-efficiency of an annotation process can be improved by reducing the amount of required manual annotation work, as we described in the previous sections. It is, however, no less important to simplify and increase efficiency of the remaining manual annotation process itself, in other words, to provide the annotators with an intuitive and efficient annotation tool.

The primary data format of the PCEDT is the Prague Markup Language (PML; Hana and Štěpánek, 2012). PML is a general XML-based format accompanied by an application framework for complex multi-layer linguistic annotations, with tools available for browsing and editing the data (tree editor TrEd;[15] Pajas and Štěpánek, 2008), for script-processing the data (a command-line tool `btred`), and for graphically oriented and powerful searching in the data (PML-TQ;[16] Pajas and Štěpánek, 2009).

The PML framework is designed to be extensible by modules (extensions) for individual treebanks, allowing to define specific data structure, manner of displaying the data, and, very importantly, corpus-specific macros for processing the data; it is actually this extensibility that makes the PML framework highly general and powerful. Thus, annotators of discourse relations in TrEd have at their disposal more than 50 discourse-related macros (revocable either from a menu or by pressing a key or a combination of keys), incl. macros for creating and deleting an arrow, assigning a discourse type and a connective to the arrow, reversing the direction of an arrow etc., but also macros for changing the way the tree is displayed, to conform with the annotator's preferences (Mírovský et al., 2010a).

Specifically for the current project, 12 new macros have been implemented, covering actions such as changing the start or target node of an arrow, transforming a projected PDTB arrow to a Prague discourse arrow, editing new Prague arrow properties (such as a PDTB sense), or confirming an automatically assigned discourse type. Some previously existing macros were redefined to not only do some action (like changing a discourse type of an arrow) but also leave a mark of such a change in the arrow attribute *comment*. The comment keeps the history of changes of an arrow; for example, the comment "`PHASE:3 RETYPED confr->synchr; INTRA_AUTO_OPPOSITE-NODES-PDTB_INCOMPATIBLE-SENSE_SIMILAR-CONN`" means that an intra-sentential AUTO arrow was merged with a PDTB arrow that shared start and target nodes (but in the opposite direction), had incompatible discourse type vs. sense, and their connectives partially overlapped; afterwards, the discourse type of the arrow was manually changed from *confrontation* to *synchrony*.

[15] https://ufal.mff.cuni.cz/tred/

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

5. Evaluation

Annotation of discourse relations is an inherently difficult task (see, e.g., Hoek and Scholman, 2017; Spooren and Degand, 2010). To evaluate the quality of the presented method, we have annotated 28 documents (1,045 sentences) completely manually and measured the inter-annotator agreement between these manually annotated documents and the documents coming from the described method, using the connective-based IAA measure presented in Mírovský et al. (2010b).[17] In our measurement, the F1 on recognition of a relation was 0.87, accuracy of discourse types assigned to relations recognized in both versions was 78% with Cohen's Kappa 0.73. If the inter-annotator agreement was measured before applying the presented method, i.e. between the automatically parsed relations (AUTO arrows) and the completely manual annotation, the results were: F1=0.84, accuracy=74%, Kappa=0.68.[18] To put these numbers in perspective, Poláková et al. (2013) reported the following numbers of inter-annotator agreement between human annotators for the first version of the Prague Discourse Treebank: F1 on recognition of a relation was 0.83, accuracy of discourse types assigned to relations recognized in both versions was 77%, with Cohen's Kappa 0.71.

6. Conclusion

We have presented a method of employing various existing language resources to induce high-quality annotation of discourse relations in a large Czech corpus, the Czech part of the Prague Czech–English Dependency Treebank (PCEDT-cs), lowering the amount of necessary interventions by human annotators. To sum up, approx. 2 thousand positions in the data were inspected manually during various stages of preparation, 6 thousand positions were checked to verify the existence of a relation and its arguments, and 8 thousand positions were inspected to check the discourse type or sense.

The new discourse annotation layer complements the other existing annotations of the PCEDT-cs texts (morphology, surface syntax, deep syntax and coreference), thus extending the set of layers of language description that are available for both sides in the parallel texts of the PCEDT.[19]

discourse type	count	percent
COMPARISON		
concession	1,169	4.0%
confrontation	1,075	3.7%
correction	346	1.2%
gradation	268	0.9%
opposition	4,109	14.2%
pragmatic contrast	31	0.1%
restrictive opposition	162	0.6%
CONTINGENCY		
condition	1,922	6.6%
explication	128	0.4%
pragmatic condition	106	0.4%
pragmatic reason–result	25	0.1%
purpose	1,525	5.3%
reason–result	2,715	9.4%
EXPANSION		
conjunction	11,119	38.5%
conjunctive alternative	226	0.8%
disjunctive alternative	213	0.7%
equivalence	92	0.3%
generalization	31	0.1%
instantiation	351	1.2%
specification	543	1.9%
TEMPORAL		
precedence–succession	1,852	6.4%
synchrony	896	3.1%
Total	28,904	100%

Table 1: Distribution of discourse types in the PCEDT-cs.

The corpus with its discourse annotation will be published under the Creative Commons Licence by the end of 2024, offering the scientific community another large shallow-discourse-annotated corpus, with almost 29 thousand explicit discourse relations; see Table 1 for the distribution of discourse types in the corpus. The data will be provided both in the Prague style of discourse annotation (i.e., on dependency trees of the tectogrammatical layer) and in the Penn style of discourse annotation (i.e., in a stand-off way on plain texts), using both Prague and Penn taxonomies of discourse types/senses, respectively.

[17] The connective-based IAA measure considers two annotations to be in agreement in recognition of a discourse relation if they both mark a relation with the same connective (or partially overlapping connectives).

[18] The relatively high numbers for the parsed relations come from the availability of manual annotation of the tectogrammatical layer.

[19] This may be the right place to note that the PDTB does not cover all documents from the PCEDT-en: many mostly short documents (150 documents, 816 sentences) from the source corpus (Wall Street Journal part of the Penn Treebank) were excluded from the original PDTB annotation and never re-introduced in the later versions. As they are a part of the PCEDT, we have included them in the annotations, processing them with the parser and then completely checked manually. Thus, 265 discourse relations have been annotated.

7. 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).

8. Bibliographical References

Ernie Chang, Muhammad Hassan Rashid, Pin-Jie Lin, Changsheng Zhao, Vera Demberg, Yangyang Shi, and Vikas Chandra. 2023. Revisiting sample size determination in natural language understanding. *arXiv preprint arXiv:2307.00374*.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondrej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, et al. 2012. Announcing Prague Czech–English Dependency Treebank 2.0. In *LREC*, pages 3153–3160.

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

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)*.

Matěj Kocián, Jakub Náplava, Daniel Štancl, and Vladimír Kadlec. 2022. Siamese BERT-based model for web search relevance ranking evaluated on a new Czech dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12369–12377.

Majid Laali and Leila Kosseim. 2017. Improving discourse relation projection to build discourse annotated corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 407–416.

Mark Lauer. 1995. How much is enough?: Data requirements for statistical NLP. *arXiv preprint cmp-lg/9509001*.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic alignment of Czech and English deep syntactic dependency trees. In *Proceedings of the Twelfth EAMT Conference*, pages 102–111, Hamburg, Germany. HITEC e.V.

Jiří Mírovský, Lucie Mladová, and Zdeněk Žabokrtský. 2010a. Annotation tool for discourse in PDT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, volume 1, pages 9–12, Beijing, China. Chinese Information Processing Society of China, Tsinghua University Press.

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

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

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

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

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

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. *Corpus Linguistics and Linguistic Theory*, 6:241–266.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024, in print. Announcing the Prague Discourse Treebank 3.0. Torino, Italy. In Proceedings of LREC-COLING 2024, European Language Resources Association.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82.

Daniel Zeman. 2004. *Parsing with a Statistical Dependency Model*. Ph.D. thesis, Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

9. Language Resource References

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].

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

Jiří Mírovský and Pavlína Synková and Lucie Poláková and Věra Kloudová and Magdaléna Rysová. 2021. *CzeDLex 1.0*. Institute of Formal and Applied Linguistics, Charles University. LINDAT/CLARIAH-CZ digital library. [link].

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 and Pavlína Jínová and Jiří Mírovský and Eva Hajičová and Anna Nedoluzhko and Radek Ocelák and Jiří Pergler and Lucie Poláková and Jana Zdeňková and Veronika Scheller and Šárka Zikánová. 2016. *Prague Discourse Treebank 2.0*. Institute of Formal and Applied Linguistics, Charles University. LINDAT/CLARIAH-CZ digital library. [link].

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].