Semantic-pragmatic Annotations in the Prague Dependency Treebank

Marie Mikulová, Eva Hajičová, Jiří Mírovský, Anja Nedoluzhko, Michal Novák, Pavlína Synková, Jan Štěpánek, Barbora Štěpánková, Jan Hajič
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
{mikulova,hajicova,mirovsky,nedoluzko,mnovak}@ufal.mff.cuni.cz
{synkova,stepanek,stepankova,hajic}@ufal.mff.cuni.cz

Abstract

We present semantic-pragmatic specification and annotation (ellipsis, coreference, bridging and discourse relations, information structure, scope of negation) in the multi-layer, genre-diversified, 3+ million-token Prague Dependency Treebank - Consolidated 2.0. While morphology and syntax work almost exclusively on sentence level, the semantic-pragmatic phenomena are often related to two or more neighbouring sentences and possibly to an extra-linguistic context. In the contribution, we describe these phenomena from both the linguistic perspective (form of expression, relation to syntax and morphology) and the cognitive perspective (relation to context, real world knowledge, as well as to the related processes such as thinking or reasoning) – classifying the possible relations between the semantic-pragmatic units into cognitively plausible, distinguishable, and human-understandable categories. We have applied our results to the corpus, by annotating it in its entirety. The resulting dataset is publicly and freely available, to serve for verification and further investigation of (not only) these phenomena.

1 Motivation

Meaning is not sentence-local. Many meaning phenomena arise from anchoring a sentence in context, encompassing both the linguistic context (neighbouring sentences) and the extra-linguistic context (knowledge of the situation and general knowledge of the world). Context is the basis for the interpretation of, in particular, the following phenomena:

- **coreference** and **bridging** relations (e.g., *When I met my wife, we* [=me and my wife] *rode a motorbike.*)
- **discourse** relations, including temporal and causal structure (e.g., *It was raining. Therefore,* [reason] *the match was cancelled.*)
- **implicit** events, arguments and circumstances (e.g., *John sent letter to Mary. Paul* [sent] [letter] *to Peggy.*)

These phenomena are often interrelated; for example, an implicit unit can be the source or target of a coreference relation; therefore, we have to be also consistent across these phenomena and make them co-exist in a coherent system, including their relation to the lower levels of language description, such as syntax and morphology.

2 Introduction

The *semantic-pragmatic*[1] phenomena going beyond strictly sentence-level meaning is now coming to the forefront of interest in the field of computational linguistics and poses serious challenges (cf. the specialized workshops on meaning representation (Lai and Wein, 2025); discourse and document-level inferences (Strube et al., 2025), coreference resolution (Ogrodniczuk et al., 2025b), understanding implicit and underspecified language (Pyatkin et al., 2024), natural language inference (Chen and Ku, 2023)). The challenge includes also the collection and annotation of relevant data for the subsequent applications and investigation of these phenomena.

In this contribution, we present semantic-pragmatic annotations in the **Prague Dependency Treebank - Consolidated 2.0** release[2] (PDT-C 2.0; Hajič et al., 2024; Mikulová et al., 2026). This language resource contains more than 3 million tokens (of Czech) manually annotated from morphology to surface and deep syntax including several types of semantic-pragmatic annotations: ellipsis resolution, coreference and bridging relations, discourse

[1] We use the term *semantic-pragmatic* to refer to a broad range of phenomena that are elsewhere also referred to as *inter-sentential*, *discourse-level*, *document-level*, or *context-based* and *knowledge-based*.

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

Findings of the Association for Computational Linguistics: ACL 2026, pages 21099–21110

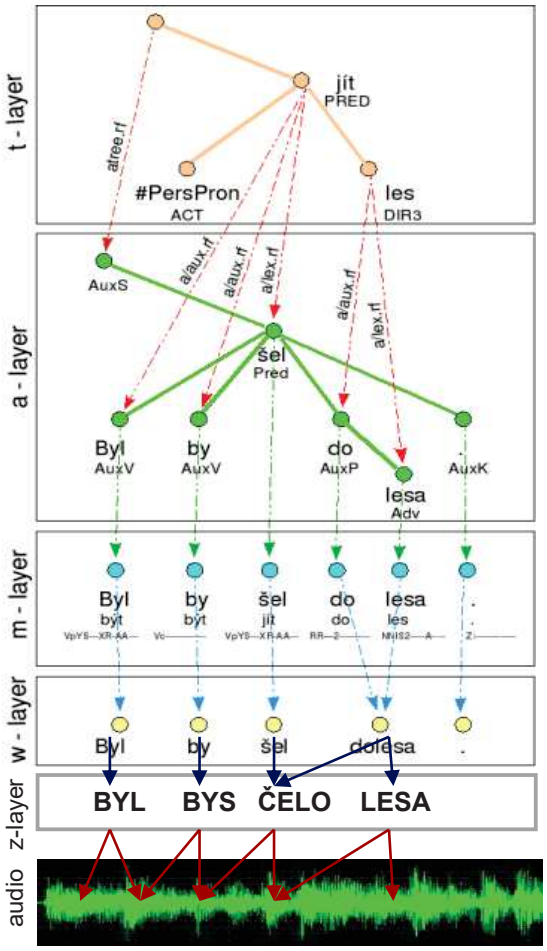

Figure 1: PDT multi-layer annotation scheme illustrated on the example of the Czech sentence: *Byl by šel do lesa.* 'lit.: He-was would went to forest.'

relations, information structure, scope of negation and focalizers. Regarding the semantic-pragmatic annotations, the corpus is unique, particularly in the following aspects:

(i) It contains multiple semantic–pragmatic phenomena annotated **simultaneously** in contrast to corpora built exclusively for a single intersentential phenomenon (see examples in Sect. 3).

(ii) The semantic–pragmatic phenomena are annotated above the **rich linguistic annotation** including morphological features, syntactic structure, and predicate-argument structure.

(iii) The specification of semantic-pragmatic phenomena is verified by manually annotating **diverse, large volumes** of textual material.

This allows the subsequent modelling and investigation to combine semantic–pragmatic information of different types (as in Tseng et al., 2021) and to follow the overall meaning-making process from semantic–pragmatic phenomena to syntax and form and vice versa (cf. Bender et al., 2015). As has been demonstrated repeatedly, incorporating different linguistic features leads to improved performance when modelling semantic–pragmatic phenomena (e.g., Khullar (2021) incorporates syntactic information in an ellipsis resolution model; He et al. (2020) investigate a similar approach in the NLI task; or Yoshida et al. (2023) used syntax-based approach in negation scope resolution). This also supports the idea of the Functional Generative Description, a well-founded linguistic theory which is behind most of the PDT work, that all levels of the language system contribute to the overall understanding, and the level of linguistic (compositional) meaning should be considered as a suitable starting point for further semantic–pragmatic interpretation (Sgall et al., 1986; Sgall, 1995).

In this contribution, we describe the semantic-pragmatic phenomena from both the linguistic perspective (form of expression, relation to syntax and morphology) and the cognitive perspective (relation to context, real world knowledge), classifying the possible relations between the semantic-pragmatic units into cognitively plausible, distinguishable, and human-understandable categories.

The paper is organized as follows: in Sect. 3, the related works (datasets annotated for semantic–pragmatic phenomena) are mentioned. The PDT annotation scheme is described in Sect. 4. The core of the paper is presented in Sect. 5, in which we describe the PDT annotation of ellipsis (Sect. 5.1), coreference (Sect. 5.2) and bridging (Sect. 5.3) relations, discourse (Sect. 5.4), information structure (Sect. 5.5), and scope of negation and focalizers (Sect. 5.6). The annotation process is briefly described in Sec. 6. We conclude in Sect. 8, describing also ongoing annotation efforts and the future work in Sect. 7.

3 Related Work

Recently, a number of survey papers have documented the shift from research on syntactic, sentence-level phenomena to semantic–pragmatic ones (cf. Ma et al., 2025; Sadeddine et al., 2024). Existing datasets are being enriched by additional semantic–pragmatic annotations (e.g., UMR adds a document-level phenomena to the AMR sentence-level annotation; Van Gysel et al., 2021), and new corpora are being developed with a primary focus on annotating semantic–pragmatic phenomena. In

the following list, we mention only the most recent ones: corpora for discourse relations: TDDiscourse (Naik et al., 2019) or eRST (Zeldes et al., 2025 – cf. a detailed list of corpora with discourse annotation there, another overview can be found in Ogrodniczuk et al., 2025a); ellipsis: NoEl (Khullar et al., 2020) and Ellipsis Corpus (Cavar et al., 2024); bridging relations: GUMBridge (Levine and Zeldes, 2026; the state of the art in bridging resolution can be found in Kobayashi and Ng, 2020); coreference: CorefUD (Novák et al., 2025 or Nedoluzhko et al., 2022 – cf. a detailed list of corpora with coreference annotation there); information structure: ThemePro (Dominguez et al., 2020); datasets for negation scope resolution are summarized in Wu and Sun (2023). To our knowledge, there are not many resources that contain different types of semantic–pragmatic annotations within a single dataset. The PDT-C 2.0 dataset is the most comparable to the aforementioned GUMBridge corpus in terms of annotation richness, as it contains annotations of information status, coreference, discourse relations, as well as syntactic and morphological annotations.

4 Prague Dependency Treebank

A manually annotated and genre-diversified language resource (more than 3 million tokens) with rich linguistic information from morphology and syntax to meaning, the Prague Dependency Treebank - Consolidated 2.0 (Hajič et al., 2024; Mikulová et al., 2026) is a consolidated release of the existing PDT-corpora of Czech data, uniformly annotated using the standard PDT multilayer scheme and enriched with several types of semantic-pragmatic annotations. It consists of four different datasets: dataset of written text, translated texts,[3] spoken texts, and of user-generated texts. The multi-layer architecture of PDT is based on the Functional Generative Description theory (Sgall et al., 1986). It is reflected in several detailed annotation manuals available from the project website[4] and here, it is schematically illustrated in Fig. 1.[5]

At the **morphological layer** (**m-layer** box in Fig. 1), a 15-character tag is used to describe morphological properties of all tokens.

At the **analytical layer** (**a-layer**), a surface syntactic structure is captured by a tree-like graph with the specification of the head for each node and the assignment of a syntactic function that corresponds to traditional syntactic categories such as subject (`Sb`), object (`Obj`), or adverbial (`Adv`).

Tectogrammatical layer (**t-layer**) captures a complex semantic-syntactic annotation of a sentence, especially its predicate-argument structure. It is also annotated for context-based, semantic-pragmatic phenomena: ellipsis, coreference, bridging, discourse relations, and information structure, which are all the subject of this paper.

In addition to these three main annotation layers, there is also the **raw text layer (w-layer)**, where the text is segmented into documents and paragraphs and the tokens are assigned unique identifiers. There is an additional **audio** and **speech transcription (z-layer)** layer(s) for spoken data.

In order not to lose any piece of the original information, tokens (nodes) at a lower layer are explicitly referred to from the corresponding closest (immediately higher) layer. These links allow for tracing every unit of annotation all the way down to the original text, or to the transcript and audio.

5 Semantic-pragmatic Annotation

In the PDT framework, semantic-pragmatic phenomena are annotated above the annotation of the semantic-syntactic (predicate-argument) structure and relations in the t-layer. The semantic-syntactic structure is captured by a dependency tree and the relations between its nodes are called *functors*; cf. the `PRED` value for predicate, `ACT` for actor, `PAT` for patient, `MEANS` for means, `DIR3` for goal in Fig. 2.[6]

In Prague Dependency Treebank - Consolidated 2.0, the following types of semantic-pragmatic phenomena are annotated:

- **Ellipsis**. Implicit (non-expressed) events and arguments are captured by means of a node added to the tree structure. In the tree-shaped graph, such a node is visualized as a square, in contrast to circular nodes, which are anchored in surface realization; cf. the added nodes for actors (`ACT`) in the first clause (with the lemma *#PersPron* (personal pronoun) for *já* 'I') as well as in the second clause (*my* 'we') in Fig. 2. See more in Sect. 5.1.

[3] Translated data is the Czech part of Prague Czech-English Dependency Treebank, consisting of the PennTreebank-Wall Street Journal (Marcus et al., 1993) translation.

[4] `https://ufal.mff.cuni.cz/pdt-c`

[5] The visualization of the t-layer annotation is simplified.

[6] The t-layer annotation contains a wide range of further information, in particular the so-called grammatemes describing semantic features such as quantity, tense, and modality. See more in Mikulová et al. (2026).

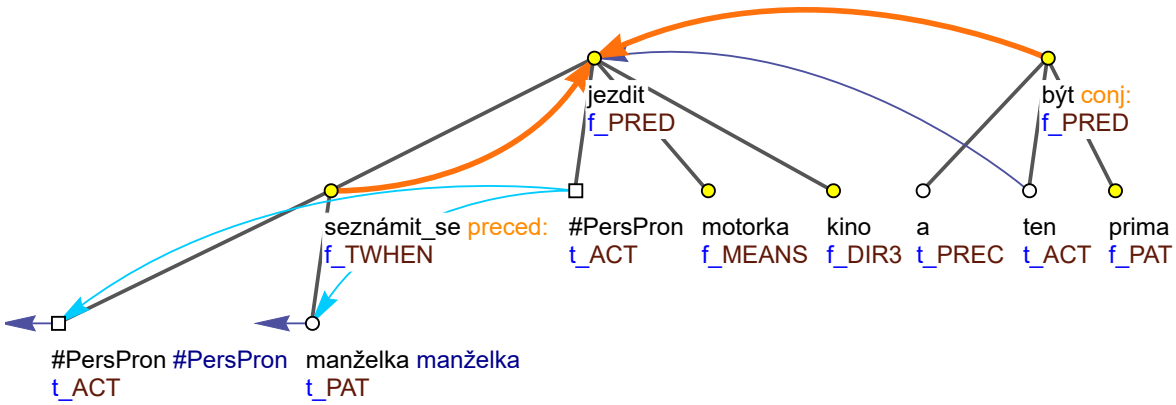

Figure 2: Semantic–pragmatic annotations in the PDT framework for the Czech sentences:
Když jsem se s manželkou seznámil, jezdili jsme na motorce do kina. A to bylo prima.
'When [I] met my wife, [we] rode our motorbike to cinema. And that was great.'

- **Coreference**. Identical coreference and non-identical association (bridging) relations between entities are represented by dark blue and light blue arrows, respectively. In Fig. 2, there is a coreference relation between the demonstrative *to/ten* 'that' in the second sentence and the information contained in the first sentence (the arrow points from the demonstrative to the predicate). Coreference is also captured for the actor (*já* 'I') and patient (*manželka* 'wife') in the first clause, with arrows from both nodes pointing to their counterparts in the preceding context (see also the dark blue antecedent lemmas depicted in the respective coreferential nodes). See Sect. 5.2.

- **Bridging**. In Fig. 2, a non-identical association (bridging) relation with a `set-subset` interpretation holds between the actor of the second clause (*my* 'we') and the actor (*já* 'I') and patient (*manželka* 'wife') of the first clause; cf. the two light blue arrows between the respective nodes.[7] See more in Sect. 5.3.

- **Discourse**. Discourse relations are depicted by the orange arrow between the predicates of the two related clauses or sentences. In Fig. 2, there is the temporal discourse relation of `preceding` event between the first and second clauses and a discourse relation of `conjunction` between the first and second sentences. See Sect. 5.4.

- **Information structure**. For each semantic unit (node), it is indicated whether it is contextually bound, or introduces new information. In Fig. 2, the blue values of `t` (preceding the functor values) mark contextually bound nodes, while the `f` labels mark contextually non-bound nodes. See more in Sect. 5.5.

- **Scope**. The scope of negation and focalizers are also annotated. For details, see Sect. 5.6.

In the following subsections, we describe the specification and annotation of individual semantic-pragmatic phenomena in more detail.

5.1 Ellipsis (Implicit Events and Arguments)

Ellipsis is a linguistic phenomenon in which parts of a sentence are omitted, although the missing parts are essential for understanding the meaning. We identify ellipsis on the basis of dependency and predicate-argument structure and/or other linguistic or extra-linguistic context. Two types of ellipses are captured: (i) **ellipsis of a parent node** (verb-parent ((1), (3))[8] or noun-parent (2)) to an orphan-dependent; and (ii) ellipsis of an obligatory dependent (**core-participant ellipsis**; (1), (4)). In both cases, a distinction is made as to whether the elided unit can be lexically reconstructed based on textual context ((1), (2)) or whether it is an unit with a general lexical value ((3), (4)).

(1) *John sent letter to Mary. But Paul* [sent] [letter] *to Peggy.* See Fig. 3.
(2) *Central* [Europe] *and Eastern Europe*
(3) [Empty Verb] *Coffee!*
(4) *the January sale* [of something] [to anyone]

[7] The relation between *we* and *I + wife* (in Fig. 2) is often understood as a *split antecedent* case of coreference.

[8] For clarity, we use English examples. The annotations in the Prague Dependency Treebank - Consolidated 2.0 are exclusively on Czech (see an example in Fig. 2).

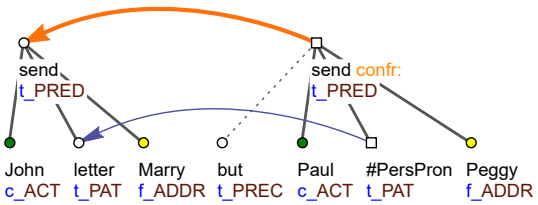

Figure 3: Semantic-pragmatic annotations in the PDT framework for the sentences: *John sent letter to Mary. But Paul to Peggy.*

All types of ellipses are captured by adding an extra node to the t-layer tree. The added parent nodes have either a lexical lemma or a general lemma (if the lexical value of the missing unit cannot be reconstructed). Context-based and general ellipses of omitted core participants (predicate arguments) are distinguished by a special lemma and there is a coreference relation (see Sect. 5.2) in the case of context-based ellipsis. An ellipsis of the event and of an obligatory participant is shown in Fig. 3, which presents a simplified t-layer annotation of (1). See more in Mikulová et al. (2006); Hajičová et al. (2015); Hajič et al. (2015).

5.2 Coreference

Coreference annotation captures a referential identity relation between **entities** ((6), (7); including **temporal** and **spatial deixis** (8)) and **partly** also between **events** (9). In annotation, a distinction is made between **grammatical** (unambiguous, intrasentential reference given by grammatical rules, such as between *himself* and *Peter* in (5)) and **context-based** coreference.

(5) *Peter only likes himself.*
(6) *We rode our motorbike to the cinema.*
(7) *Sony has long history. The firm was founded in 1946.*
(8) *He arrived in Prague. He settled there.*
(9) *We rode motorbike to the cinema. That was great.*
(10) [point to photo] *This is my fiancée.*
(11) *The debate on finance reform will begin on Wednesday. All discussions will take place behind closed doors. The minister announced this yesterday.*

Within the context-based coreference, two types are distinguished. **Specific** coreference relation holds between the same instances of an entity or event (e.g., the relation between *that* and 'riding a motorbike to the cinema' in (9) (cf. also Fig. 2), relation between *we* and *our* in (6) or *Sony* and *the firm* in (7)). A **generic** coreference is captured between the same types of entity or event (e.g., the relation between *letters* in first and second clause in Fig. 3 is a generic reference). Coreference annotation follows the "chain principle": coreferential unit always refers to the last preceding antecedent. Two more special cases of (co)reference are further marked by a special value. An **exophoric** relation is marked in the case of a reference to situational context or reality external to the text (10). A reference to a **segment** of text is marked in case of a reference to an antecedent consisting of more than one sentence (11). See more in Nedoluzhko and Mírovský (2011); Nedoluzhko et al. (2016).

5.3 Bridging

Apart from the identical coreference relations, non-identical association (bridging) relations are annotated if they are related in one of the specific types of semantic or conceptual ways to their antecedents. Several types are distinguished, esp. metonymical relation between a part and a whole (`part-of`; *room – ceiling*); relation between a set and its subsets (`set-subset`, *students – some students – a student*, or the relation between *we* and *I* with *wife* in (12); cf. also in Fig. 2), relation between an entity and a singular function on this entity (`function`; *prime minister – government*), etc.; see Tab. 1 for a complete list. There is also an `other` category for further underspecified relations whose delimitation is the subject of further research; this category includes: family (*father – son*), place – inhabitant, author – work (13), object – owner, object – typical instrument (*tightrope walker – rope*), etc. See more in Zikánová et al. (2015).

(12) *When I met my wife, we were riding a motorbike.*
(13) *The author will certainly play a large role in the selection of the painting.*

5.4 Discourse

Annotation of discourse relations (Poláková et al., 2012; Zikánová et al., 2015, 2019) was inspired by lexical approach applied in Penn Discourse Treebank (Prasad et al., 2008). It covers explicit local relations marked by discourse connectives that hold between two spans of text (usually clauses and sentences) called discourse arguments. These arguments occur either within one sentence, or in different sentences. Annotation includes both grammaticalized mostly one-word connectives (so-called primary connectives) such as *a* 'and', *ale* 'but', *když* 'when' or *protože* 'because', and more complex expressions with connective function (secondary connectives) such as *z toho důvodu* 'for that reason' or *to znamená, že* 'it means that' which are not fully grammaticalized and form an open class (Rysová and Rysová, 2015).

Discourse relations are annotated between roots of the relevant subtrees (sentences, clauses, phrases) in the trees and depicted by orange arrows oriented and labelled according to individual discourse types. In PDT-C 2.0, 22 discourse types such as `reason-result` (14), `condition`, `purpose`, `equivalence`, etc. are distinguished; see Tab. 1 for a complete list. In Fig. 2, there are two discourse relations: `precedence-succession` relation (15), and `conjunction` relation (16).

As `precedence-succession` is an asymmetric relation (the semantics of the arguments differ), it is oriented from the previous event "meeting the wife") to the next one ("riding to the cinema"). On the other hand, `conjunction` represents a symmetrical relation (a relation in which the roles of the arguments do not differ), so the arrow representing the relation is oriented from right to left (as for all symmetric relations).

(14) *It was raining. Therefore, the match was cancelled.*
(15) *When I met my wife, we rode a motorbike.*
(16) *When I met my wife, we rode a motorbike. And that was great.*

Many pieces of information have been added to each discourse relation in the data: namely the (lexical) connective, information on the extent of the argument,[9] information whether the connective is primary or secondary, etc.

Discourse type taxonomy was inspired by the Penn Discourse Treebank and the Czech syntactic tradition. For better usability of annotation in the international scientific community, all discourse types are also transformed into PDTB 3.0 labels (Webber et al., 2019).[10] In Prague Dependency Treebank - Consolidated 2.0, the discourse relations are labelled according to both taxonomies. E.g., the discourse relation in Fig. 3 has PDT type `confrontation` and also corresponding PDTB 3.0 sense *Comparison.Contrast*. All PDT labels are listed in Tab. 2 in the Appendix A together with the corresponding PDTB 3.0 labels.[11]

[9]Typically, the extent of an argument corresponds to a subtree of the source or target node of the arrow (with the exclusion of the other argument in subordinated constructions). A group of range attributes also allow to include one or more subsequent sentences, or even define an arbitrary set of nodes representing the argument.

[10]For details on transformation, see Mírovský et al. (2023).

[11]The discourse annotation in the PDT-C 2.0 is also converted into the Penn Discourse Treebank format within the Prague Discourse Treebank 4.0 (Synková et al., 2024; Mírovský and Synková, 2026).

5.5 Information Structure

Our model of the information structure[12] is based on the notion of "communicative dynamism" combined with the "aboutness" relation. The concept of "aboutness" splits the sentence (or more precisely, its meaning representation) into "what the utterance is about" (topic) and "what is being said about it" (focus). Furthermore, the PDT framework distinguishes between the actual linear arrangement of words (surface structure at a-layer) and the interpretive arrangement based on increasing "communicative dynamism" (deep structure at t-layer). This helps to analyze the relation between word order in "free-word order" languages like Czech and the communicative strategy. It is also prominent in distinguishing between spoken and written text, with spoken language often using prosody (stress, intonation), instead of, or together with, word order to signal communicative dynamism. In the actual annotation scheme, information structure is captured by the fundamental dichotomy "given-new" as both an opposition and a scale:

- **Contextual boundness**. For each semantic unit (i.e., each t-layer node), it is determined whether the information it represents is already known from the context or is new. In general, for each semantic unit is marked whether it belongs to the topic or to the focus.

- **Communicative dynamism**, i.e., an ordering of semantic units from the least communicatively salient elements of the sentence to the elements with the highest degree of "communicative salience" (Hajičová and Sgall, 2001).

In the annotation of contextual boundness, three values are distinguished: `t` (topic) denotes contextually bound (given) nodes, `f` (focus) is used for contextually unbound (new) nodes, and `c` (contrastive) is used for contrastively contextually bound nodes. Contrastive contextual boundness (value `c`) can be based, for example, on a bridging relation (see Sect. 5.3): the annotation of contextual boundness in Fig. 3 corresponds to the context indicated in example (17), *John* and *Paul* are contrastively bound units based on the bridging relation `subset-set`.

[12]This phenomenon is referred to by various terms, such as *theme-rheme*, *topic-focus articulation*, *thematic progression*, *functional sentence perspective*. We use the general term of *information structure*. Arguments that this phenomenon is semantically relevant and significantly contributes to the study of the functioning of language can be found in Hajičová (2019) and were previously noted by Halliday (1967).

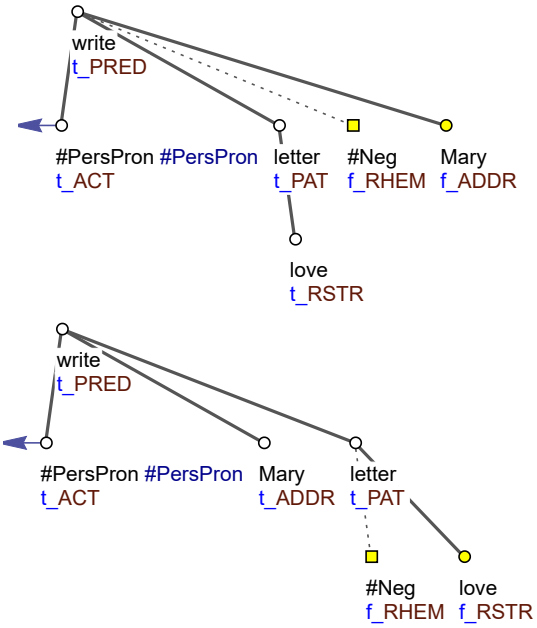

Figure 4: Two different PDT framework annotations for the sentence: *He didn't write a love letter to Mary.* The upper one corresponds to this sentence in the context "not to Mary, but to Susan", the bottom one is for the context "not love, but business".

(17) [The brothers wrote letters.] *John sent letter to Mary. But Paul to Peggy.*

Communicative dynamism is represented by the node (re)ordering of the t-layer tree. The degree of communicative dynamism of a particular node is determined relative to its parent and sister nodes, that is, for each level of the tree. The rightmost node in the (sub)tree is the node with the highest degree of communicative dynamism. For example, in the sentences represented in Fig. 3, the nodes representing the addressees of the letter-sending event (*to Mary*, *to Peggy*) have the highest degree of communicative dynamism.

5.6 Scope of Negation and Focalizers

Through the arrangement of nodes in the tree, we also capture the scope of negation and so-called focalizers, i.e., expressions whose function is to signal the information structure of a sentence by highlighting the most communicatively salient elements, namely focus ((19)–(22)) and contrastive topic (18). This is important for understanding the meaning of the sentence as a whole, and it offers insight into the interaction between semantic structure and world knowledge; see, for example, recent research in NLP (Kamath et al., 2024).

Negation markers and focalizers do not always precede, in surface word order, the units they negate or emphasize – that is, the elements within their scope; cf. (19)-(22), in which (depending on the context) a different part of the sentence falls within the scope of negation.

Annotation	Count
ELLIPSIS	
core-participant	223,409
noun parent	5,928
empty noun parent	1,530
verb parent	12,809
empty verb parent	9,889
DISCOURSE RELATIONS: TOTAL	82,506
Discourse relations: COMPARISON	
concession	2,736
confrontation	2,432
correction	1,390
gradation	1,108
opposition	10,348
pragmatic contrast	231
restrictive opposition	1,019
Discourse relations: CONTINGENCY	
condition	4,801
explication	373
pragmatic condition	372
pragmatic reason-result	493
purpose	2,710
reason–result	10,927
Discourse relations: EXPANSION	
conjunction	29,509
conjunctive alternative	854
disjunctive alternative	623
equivalence	544
generalization	522
instantiation	687
specification	1,428
Discourse relations: TEMPORAL	
precedence–succession	6,600
synchrony	2,799
COREFERENCE AND BRIDGING	
grammatical coreference	66,168
textual coreference	331,783
bridging anaphora: subset-set	29,308
bridging anaphora: part of	6,614
bridging anaphora: contrast	2,318
bridging anaphora: function	2,292
bridging anaphora: non-coref. anaphora	851
bridging anaphora: other	2,252
segment of text	5,903
situational context	20,753

Table 1: Volume of semantic-pragmatic annotations in Prague Dependency Treebank - Consolidated 2.0.

(18) [The brothers were out.] *Only Charlie went home.*
(19) *He didn't write love letter to Mary.* [but went out]
(20) *He didn't write love letter to Mary.* [but book]
(21) *He didn't write love letter to Mary.* [but to Susan]
(22) *He didn't write love letter to Mary.* [but business]

In the PDT framework, the scope of negation and focalization is captured by the ordering of nodes in the tree. The node representing a scope-bearing expression is placed as the closest left sister of the expression that falls within its scope.[13] See the different node orderings in the representations of sentences (21) and (22), including the different placement of the negation node (which has the special lemma *#Neg*) in Fig. 4. Further information on the scope annotation can be found in Mikulová et al. (2006).

6 Annotation Process

All of the above-described semantic-pragmatic phenomena are annotated throughout the entire Prague Dependency Treebank - Consolidated 2.0, with the exception of bridging relations (Sect. 5.3), information structure (Sect. 5.5), and scope annotations (Sect. 5.6), which are present only in the written-text part (833,180 tokens) of the PDT-C 2.0 dataset (cf. Mikulová et al., 2026). The volume of semantic-pragmatic annotations in PDT-C 2.0 is presented in Tab. 1.

All types of annotation were performed manually using different methods. Annotations of ellipsis, coreference[14] and bridging relations, information structure, and scope were carried out completely manually. Discourse annotation was completed using different methods: the written-text part of the dataset was annotated completely manually; the translated part was annotated using annotation projection from Penn Discourse Treebank and a parser based on previously manually annotated written-text part of the dataset, both types of relations (obtained via projection and parser) were merged or modified manually reaching quality comparable to fully manual annotation. In the spoken part, automatic pre-annotation was followed by thorough manual checks reaching again high quality compared to fully manual annotation, and the user-generated part of the dataset was annotated by the parser and completely manually checked.

The annotation process for a particular type of semantic-pragmatic phenomenon, along with information on inter-annotator agreement, has been thoroughly documented in recently published or submitted papers and technical reports (summarized in Mikulová et al., 2013; Mikulová, 2014; Mikulová et al., 2022; discourse annotation: Mírovský and Synková, 2026; coreference and bridging: Nedoluzhko and Mírovský, 2011; Nedoluzhko et al., 2011, 2016; information structure and scope: Veselá et al., 2004; Mírovský et al., 2013).

[13] If the parent node is within the scope, the scope-bearing expression is placed as the rightmost left direct child of the parent node.

[14] Grammatical coreference in the translated, spoken, and user-generated datasets was automatically preprocessed and then fully manually checked.

7 Ongoing Annotation and Future Work

The description of the language system is far from complete. Fundamental linguistic research thus remains essential, including the relation of language to the real world, in its semantic-pragmatic aspects. In this respect, we will continue our efforts in systematically describing language from form to meaning and beyond. For some of the phenomena, we can rely on previous theoretical work, adding extensions and adaptations that come with the projection of the theoretical results to actual text annotation. The following annotation efforts have already been initiated: (i) event-type annotation (Urešová et al., 2025), (ii) implicit discourse relations (Zikánová et al., 2019), and (iii) fine-grained classification of (circumstantial) semantic roles, such as temporal, spatial, and manner (Mikulová, 2024).

We will also consider other semantic-pragmatic phenomena that are currently not included in the PDT ecosystem. For example, there has been recent work on event coreference, scope of quantification, or (global) temporal and modal structure (e.g., Van Gysel et al., 2021). These phenomena will also be considered for future specification consistent with the current PDT-based approach, and then annotated in future versions of the Prague Dependency Treebank - Consolidated.

8 Conclusions

We presented a dataset containing interrelated semantic-pragmatic annotations (e.g., ellipsis, coreference, bridging, discourse relations, information structure and scope of negation). In the publicly available and freely downloadable Prague Dependency Treebank - Consolidated 2.0, we are providing all these phenomena manually annotated over real texts and integrated with the original annotation, from morphology and syntax to meaning and the semantic-pragmatic phenomena described in this contribution.

In the context of the current discussion about the importance or non-importance of the work on "traditional" issues in computational linguistics, we

believe that pursuing the *whys* and *hows* regarding linguistic phenomena and their relation to pragmatics and the real world knowledge is as important as it was when computational linguistics had been conceived as a field of science. First of all, we are interested in the language system and its relation(s) to the real world, our knowledge of it, as well as the related processes such as thinking or reasoning. We also believe this is independent of the advances that we are experiencing in the area of AI and Large Language Models. While LLM's behavior is undoubtedly improving and will continue to do so, and even if their performance in various applications cannot be distinguished (from the outside observation) from humans, or is even better, it still makes sense to investigate language as a system and its relation to the real world. We hope that our contribution to the description of such a relation, which we have presented in this paper and in the annotated data and its specification and detailed description, is a step towards this goal.

9 Limitations

There are several limitations we are aware of. Some of the phenomena presented here (bridging relations, information structure, scope of negation) are not annotated across the entire corpus, but only in part of it (even so, it still represents a large volume of annotation).

We are also aware of the fact that the annotation described here is carried out on Czech texts only. However, the formal specification and description of the phenomena have been prepared with other languages in mind, especially English (one part of Prague Dependency Treebank - Consolidated 2.0 is a translation from English; therefore, we have the source available for at least bilingual verification).

Acknowledgments

The work described herein has been supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project LM2023062 (LINDAT/CLARIAH-CZ)[15], project LUAUS23283 (Uniform Meaning Representation), and by Charles University Research Centre program No. 24/SSH/009 (Multilingual Lens: Investigating Large Text Corpora from Different Methodological Perspectives).

[15] https://lindat.cz

References

Emily M. Bender, Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of Interpretation: On Grammar and Compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London, UK. Association for Computational Linguistics.

Damir Cavar, Ludovic Mompelat, and Muhammad Abdo. 2024. The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 46–54, St. Julian's, Malta. Association for Computational Linguistics.

Kehai Chen and Lun-Wei Ku, editors. 2023. *Proceedings of the Second Workshop on Natural Language Interfaces*. Association for Computational Linguistics, Bali, Indonesia.

Monica Dominguez, Juan Soler, and Leo Wanner. 2020. ThemePro: A Toolkit for the Analysis of Thematic Progression. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1000–1007, Marseille, France. European Language Resources Association.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2024. Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, http://hdl.handle.net/11234/1-5813.

Jan Hajič, Eva Hajičová, Marie Mikulová, Jiří Mírovský, Jarmila Panevová, and Daniel Zeman. 2015. Deletions and Node Reconstructions in a Dependency-Based Multilevel Annotation Scheme. In *Computational Linguistics and Intelligent Text Processing*, pages 17–31, Cham. Springer.

Eva Hajičová. 2019. A Plea for Information Structure as a Part of Meaning Representation. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 66–72, Florence, Italy. Association for Computational Linguistics.

Eva Hajičová, Marie Mikulová, and Jarmila Panevová. 2015. Reconstructions of Deletions in a Dependency-based Description of Czech: Selected Issues. In

Proceedings of the Third International Conference on Dependency Linguistics, pages 131–140, Uppsala, Sweden.

Eva Hajičová and Petr Sgall. 2001. Topic-focus and Salience. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 276–281, Toulouse, France. Association for Computational Linguistics.

M. A. K. Halliday. 1967. *Intonation and Grammar in British English*. Mouton, The Hague–Paris.

Qi He, Han Wang, and Yue Zhang. 2020. Enhancing Generalization in Natural Language Inference by Syntax. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4973–4978, Online. Association for Computational Linguistics.

Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. Scope Ambiguities in Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:738–754.

Payal Khullar. 2021. Are Ellipses Important for Machine Translation? *Computational Linguistics*, 47(4):927–937.

Payal Khullar, Kushal Majmundar, and Manish Shrivastava. 2020. NoEl: An Annotated Corpus for Noun Ellipsis in English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 34–43, Marseille, France. European Language Resources Association.

Hideo Kobayashi and Vincent Ng. 2020. Bridging Resolution: A Survey of the State of the Art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kenneth Lai and Shira Wein, editors. 2025. *Proceedings of the Sixth International Workshop on Designing Meaning Representations*. Association for Computational Lingustics, Prague, Czech Republic.

Lauren Levine and Amir Zeldes. 2026. GUMBridge: a Corpus for Varieties of Bridging Anaphora. In *Proceedings of the 15th Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the Era of Large Language Models: A Survey on Datasets, Evaluation, Opportunities and Challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marie Mikulová. 2014. Annotation on the Tectogrammatical Level. Additions to Annotation Manual (with respect to PDTSC and PCEDT). Technical Report TR-2013-52, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Marie Mikulová. 2024. Fine-grained Classification of Circumstantial Meanings within the Prague Dependency Treebank Annotation Scheme. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 7314–7323, Torino, Italia.

Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, and Zdeněk Žabokrtský. 2013. From PDT 2.0 to PDT 3.0 (Modifications and Complements). Technical Report TR-2013-54, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Marie Mikulová, 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. 2006. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation manual. Technical Report TR-2006-30, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Marie Mikulová, Jiří Mírovský, Milan Straka, Pavlína Synková, Barbora Štěpánková, Jan Štěpánek, and Jan Hajič. 2026. Prague Dependency Treebank - Consolidated 2.0: Enriching a Complex Annotation Scheme. In *Proceedings of the 15th Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. 2022. Quality and Efficiency of Manual Annotation: Pre-annotation Bias. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France. European Language Resources Association.

Jiří Mírovský, Kateřina Rysová, Magdaléna Rysová, and Eva Hajičová. 2013. (Pre-)Annotation of Topic-Focus Articulation in Prague Czech-English Dependency Treebank. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 55–63, Nagoya, Japan. Asian Federation of Natural Language Processing.

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 and Pavlína Synková. 2026. Presenting the Prague Discourse Treebank 4.0. In *Proceedings of the 15th Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A Dataset for Discourse-Level Temporal Ordering of Events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, and Lenka Šíková. 2011. Annotation of Coreference in Prague Czech-English Dependency Treebank. Technical Report TR-2014-57, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Anna Nedoluzhko and Jiří Mírovský. 2011. Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank. Technical Report TR-2011-44, Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 169–176, Portorož, Slovenia. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference Meets Universal Dependencies. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2025. Findings of the Fourth Shared Task on Multilingual Coreference Resolution: Can LLMs Dethrone Traditional Approaches? In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 95–118, Suzhou, China. Association for Computational Linguistics.

Maciej Ogrodniczuk, Anna Latusek, Karolina Saputa, Alina Wróblewska, Daniel Ziembicki, Bartosz Żuk, Martyna Lewandowska, Adam Okrasiński, Paulina Rosalska, Anna Śliwicka, Aleksandra Tomaszewska, and Sebastian Żurowski. 2025a. Where Frameworks (Dis)agree: A Study of Discourse Segmentation. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences*, pages 182–196, Suzhou, China. Association for Computational Linguistics.

Maciej Ogrodniczuk, Michal Novák, Massimo Poesio, Sameer Pradhan, and Vincent Ng, editors. 2025b. *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Suzhou, China.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Valentina Pyatkin, Daniel Fried, Elias Stengel-Eskin, Alisa Liu, and Sandro Pezzelle, editors. 2024. *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*. Association for Computational Linguistics, Malta.

Magdaléna Rysová and Kateřina Rysová. 2015. Secondary Connectives in the Prague Dependency Treebank. In *Proceedings of the Third International Conference on Dependency Linguistics*, pages 291–299, Uppsala, Sweden. Uppsala University.

Zacchary Sadeddine, Juri Opitz, and Fabian Suchanek. 2024. A Survey of Meaning Representations – From Theory to Practical Utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2877–2892, Mexico City, Mexico. Association for Computational Linguistics.

Petr Sgall. 1995. From Meaning via Reference to Content. In *Karlovy Vary studies in reference and meaning*, pages 172–183. Filosofia Publications, Prague, Czech Republic.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague/Dordrecht.

Michael Strube, Chloe Braud, Christian Hardmeier, Junyi Jessy Li, Sharid Loaiciga, Amir Zeldes, and Chuyuan Li, editors. 2025. *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences*. Association for Computational Linguistics, Suzhou, China.

Pavlína Synková, Jiří Mírovský, Marie Paclíková, Lucie Poláková, Magdaléna Rysová, Veronika Scheller, Jana Zdeňková, Šárka Zikánová, and Eva Hajičová. 2024. Prague Discourse Treebank 4.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, http://hdl.handle.net/11234/1-5680.

Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin

Li, and Hong Yu. 2021. CREAD: Combined Resolution of Ellipses and Anaphora in Dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3390–3406, Online. Association for Computational Linguistics.

Zdeňka Urešová, Eva Fučíková, Cristina Fernández Alcaina, and Jan Hajič. 2025. Linking an Event-type Ontology to Morphosyntax of the Predicate-Argument Structure. *Dictionaries: Journal of the Dictionary Society of North America*, 46(1):207–227.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *KI-Künstliche Intelligenz*, 35(3-4):343–360.

Kateřina Veselá, Jiří Havelka, and Eva Hajičová. 2004. Annotators' Agreement: The Case of Topic-Focus Articulation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal. European Language Resources Association.

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

Yin Wu and Aixin Sun. 2023. Negation Scope Refinement via Boundary Shift Loss. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6090–6099, Toronto, Canada. Association for Computational Linguistics.

Asahi Yoshida, Yoshihide Kato, and Shigeki Matsubara. 2023. Revisiting Syntax-Based Approach in Negation Scope Resolution. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 18–23, Toronto, Canada. Association for Computational Linguistics.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics*, 51(1):23–72.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Charles University, Prague, Czech Republic.

Šárka Zikánová, Jiří Mírovský, and Pavlína Synková. 2019. Explicit and Implicit Discourse Relations in the Prague Discourse Treebank. In *Proceedings of the 22nd International Conference on Text, Speech and Dialogue*, volume 11697 of *Lecture Notes in Computer Science*, pages 236–248, Cham / Heidelberg / New York / Dordrecht / London. University of West Bohemia, Springer International Publishing.

A Appendix

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

Table 2: Basic transformation from PDT-C 2.0 discourse types to the PDTB 3.0 second-level senses (third-level are covered by orientation of the PDT relations)