DReUD: Discourse Relations in Universal Dependencies

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

Abstract

We present a proposal for an annotation scheme and data representation of shallow discourse relations annotation in the Universal Dependencies (UD) framework, as a theoretically appropriate and also practically oriented extension of the established morphosyntactic analysis. We outline the design requirements for the annotation scheme, encompassing simplicity, comprehensibility, theoretical grounding, practical applicability and technical robustness, while accommodating the specific constraints of shallow discourse analysis. At the same time, we present a work-in-progress baseline version of DReUD (Discourse Relations in Universal Dependencies), a modular shallow discourse parser for Universal Dependencies as a command-line program, a web client and a REST API service for Czech and English, designed for a seamless and rapid integration of discourse relations analysis both in the theoretical research and in NLP applications.

Keywords: Universal Dependencies, discourse relations, CoNLL-U data format, discourse parsing, web service

1. Introduction

Discourse relations are semantico-pragmatic relations between text units that convey a specific meaning such as causality, conjunction, contrast or generalization. Discourse relations anchored with a connective cue are generally called explicit, whereas those without a connective are called implicit. The following Czech example represents two text units (clauses) in a relation of *reason–result*, signalled in this case by the connective *tak* [*so*]:

(1) *Bolí mě nohy*, tak **zůstanu doma**.
[*My feet hurt*, so **I'm staying home**.]

Discourse relations are not only an integral part of the theoretical description of language but they also play a pivotal role in practical applications, as their automatic recognition (discourse parsing) is essential for a variety of beyond-sentence-level NLP tasks. In text summarization, discourse relations help identify key sentences that capture the main ideas, improving the coherence of summaries (Xu et al., 2019; Pu et al., 2023). In machine translation, preserving discourse relations ensures that the logical flow of the source text is maintained in the target language, enhancing translation quality and coherence (Joty et al., 2017; Smith and Specia, 2018). Question answering systems benefit from discourse annotation by locating relevant contexts across multiple sentences, improving response accuracy (Verberne et al., 2007). In sentiment analysis, discourse relations provide context for interpreting emotions, disambiguating sentiment in complex texts (Mukherjee and Bhattacharyya, 2012).

From many approaches to the discourse relations analysis, two have become the most prominent in corpora annotation. While Rhetorical Structure Theory (RST; Mann and Thompson, 1988) considers the text as a sequence of minimal units that are all connected by relations recursively creating a single hierarchically organized structure for the whole document, the lexical approach applied in the Penn Discourse Treebank annotation (PDTB; Prasad et al., 2008) focuses on connectives as anchors for local relations, not making any postulates on the type of a higher-level structure created by these relations (therefore it is called shallow).

Within the theoretical framework of Functional Generative Description (Sgall, 1967; Sgall et al., 1986), applied in the family of Prague Dependency Treebanks (Hajič et al., 2024), discourse relations analysis follows the PDTB approach and, at the same time, is conceptualized as an overlay upon the syntactic parsing of a text (Zikánová et al., 2015). This stems from the fact that the smallest elements linked by these relations typically consist of clauses and sentences—entities that are precisely delineated within syntactic dependency trees and recognized through the parsing process. This advantage seems to be language-independent: syntactic parsing helps locate and delineate units connected by discourse relations and identify the connectives.

Universal Dependencies (UD) is a universally recognized and widely used framework for morphosyntactic annotation of human languages that provides a consistent and cross-linguistically applicable set of grammatical relations and dependency labels, and allows for annotation of corpora of practically any language in a standardized way. This framework has been used for the morphosyntactic annotation of more than 100 languages (De Marn-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6641–6646
11–16 May 2026.

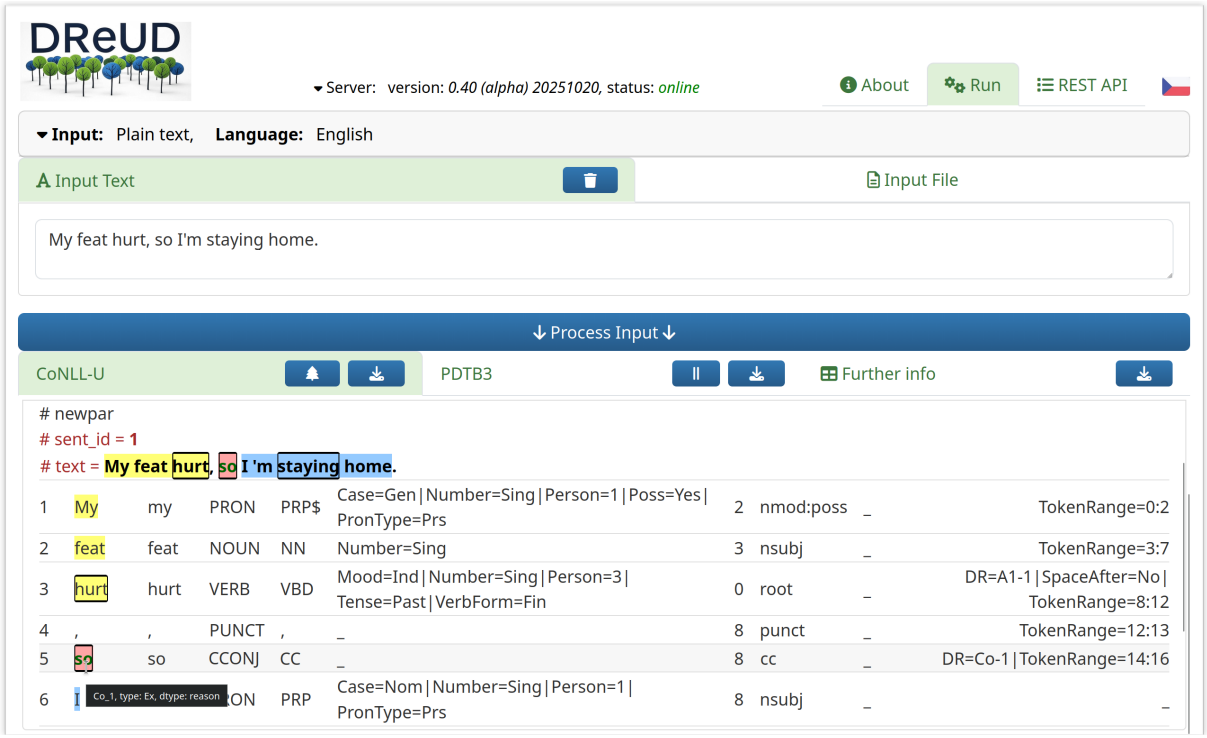


Figure 1: DReUD web interface with a (part of a) discourse relation highlighted in the CoNLL-U format

effe et al., 2021).

In recent years, the data represented in Universal Dependencies have been used not only for tasks within the sentence boundary but also for text-related tasks such as information extraction (Atmani and Lafourcade, 2021) or sentiment analysis (Kanayama and Iwamoto, 2020). The Coreference in Universal Dependencies initiative (CorefUD; Nedoluzhko et al., 2022) collected coreference corpora in different languages and harmonized them to the same scheme in the data format of Universal Dependencies (CoNLL-U), demonstrating the usefulness of including text relations in data annotated according to the UD framework.

To the best of our knowledge, the only UD language resource that includes discourse relations is the UD transformation of the GUM corpus (Zeldes, 2017). Their approach, naturally, fits the specific needs of RST corpora. An example of an annotation scheme for representing shallow discourse relations in UD was given in Pribytkova (2025). Although it is a plausible approach, it fails to fit several requirements that we enumerate in the next section.

The present paper aims at bringing annotation of shallow discourse relations into the UD framework. Our goal is two-fold: (i) to design a suitable theoretical scheme and data format for shallow discourse relations analysis in UD, and (ii) to allow for rapid usability of shallow discourse relations parsing within the UD framework both in theoretical research and practical NLP applications, by implementing such a parser as a REST API server with a practical web client.

In Section 2, we propose an annotation scheme and data format for capturing shallow discourse relations in UD. Section 3 presents an interface of the DReUD application. The architecture of the baseline parser under development used in DReUD is described in Section 4. We conclude and outline the future plans in Section 5.

2. Annotation Scheme

An annotation scheme for shallow discourse relations in Universal Dependencies and its data representation in the CoNLL-U format should comply with a number of requirements: it should be (i) simple and comprehensible, (ii) theoretically appropriate, and (iii) practical. As we are dealing with shallow discourse analysis, it should be also based on assumptions that (iv) arguments of the relations may not be immediately adjacent, (v) we do not aim at explicitly capturing higher levels of text hierarchy, and, (vi) although at the present moment we are dealing with explicit relations only, it should be suitable also for implicit relations.[1] From the technical point of view, it (vii) should be robust enough to allow, for example, for splitting a document or correcting sentence segmentation.

Our proposal is inspired both by the approach to discourse relation annotation used in the Prague

[1] and possibly other types of relations, such as EntRels (entity-based relations) from the PDTB framework

Discourse Treebank (Mírovský and Synková, 2026), (Synková et al., 2024) and in the Penn Discourse Treebank (Prasad et al., 2019), and adapted to the Universal Dependencies framework:

- Each of the two arguments of a discourse relation is represented by the root node of the subtree in the UD dependency tree that corresponds to the argument.
- For relations in subordinating constructions, the structurally embedded argument is implicitly understood to be excluded from the outer argument.
- Given the specific technical solution of coordinations in UD, also coordinated nodes and their subtrees are automatically excluded from the argument (unless included explicitly, see below).
- If an argument consists of a coordination of clauses, all roots of the coordinated clauses are explicitly marked as (a part of) the argument.
- The left argument in coordinated structures or in inter-sentential relations, or the governing one in subordinated structures is called Argument 1. The other argument is Argument 2.
- Each node belonging to a connective is marked as such and excluded from the arguments.
- Each discourse relation has a unique identifier (a number), through which the individual pieces of information of a single discourse relation are interlinked.
- Information about the type of the relation (e.g., *Explicit*, *Implicit*), the discourse type (e.g., *reason–result*),[2] and potentially any other information attached to the relation is marked at Argument 2.

A technical solution for capturing the information about a discourse relation in the *misc* column of the CoNLL-U format is exemplified in the following section.

3. DReUD - Web Interface

One of key objectives of DReUD is to allow for a simple and rapid usability, thus calling (beyond a mere command-line utility) for an intuitive web interface with multiple output formats, alongside a REST API service for seamless integration into a potentially wide range of external applications. Accordingly, DReUD delivers the output not only in CoNLL-U but also in the Penn Discourse Treebank (PDTB) format, the latter being widely used in the plain-text shallow discourse analysis.

Figure 1 shows the DReUD interface with the recognized discourse relation in the sentence from

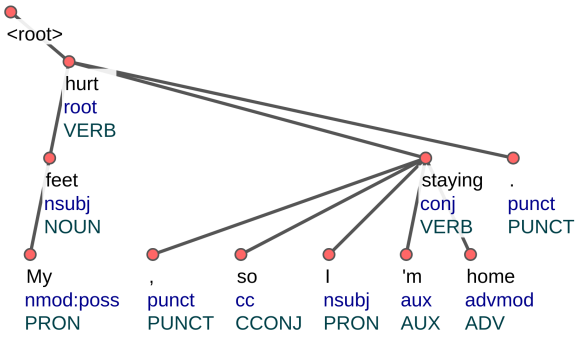
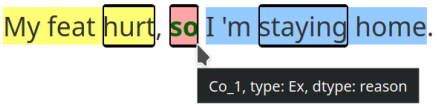


Figure 2: A dependency tree of the example sentence parsed by UDPipe

Example 1 displayed in the CoNLL-U format tab. When hovering over the connective *so*, the text spans and individual tokens corresponding to the arguments of the given relation are highlighted,[3] the roots of the syntactic subtrees of the arguments are enclosed in boxes (*hurt* and *staying*) and the classification of the relation is given in a tooltip (*Explicit*, discourse type *reason–result*).

The information is encoded in the *misc* column of the CoNLL-U format: the connective is marked with `DR=Co-1`, the root node of Argument 1 with `DR=A1-1`. The root of Argument 2 is marked with `DR=Ex-1-reason` (outside the scope of Figure 1), providing information about the type (*Explicit*) and discourse type (*reason–result*) of the relation. The three members of the relation (the connective and the roots of the arguments) are interlinked via the index `1`.

In the PDTB tab of the interface, the plain input text is displayed with only the connectives highlighted. Hovering over a connective triggers the same colouring of the members of the given relation as in the CoNLL-U tab:



Switching to the underlying format, the discourse relations in the document are displayed in the PDTB stand-off column format; in our simple example, it consists of a single entry:

```
Explicit|14..16||||||so|reason||||||
0..12||||||21..33||||||||||||14..16|
DReUD|1|reason|so|||My feat hurt||I'm
staying home|||
```

At the present moment, DReUD only uses the Prague taxonomy of discourse relations. However, support for the Penn senses is planned. As was shown in Synková et al. (2024), the Prague dis-

[2] Please note that at the general level, this proposal does not presuppose a specific taxonomy of discourse relations.

[3] using colours similar to the PDTB Annotator tool (Lee et al., 2016)

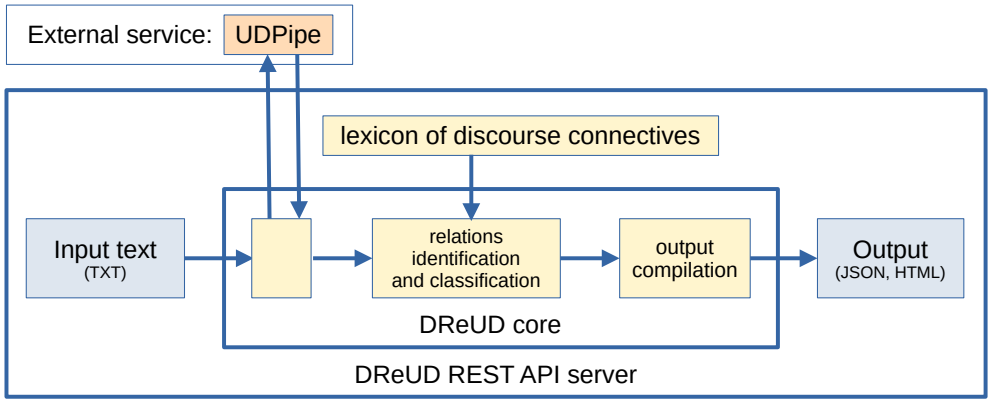


Figure 3: DReUD web interface and REST API server architecture

course types can be automatically transformed to the Penn senses with high degree of reliability.

4. DReUD – A Baseline Parser

At the core of the DReUD application, there is a modular, at the present development stage work-in-progress baseline shallow discourse parser for explicit discourse relations in Czech and English.

Figure 3 shows the architecture of the system. First, the input plain text is processed via an external web service UDPipe version 2 (an earlier iteration of which is detailed in Straka, 2018).[4] UDPipe serves as a comprehensive pipeline for sentence segmentation, tokenization, lemmatization, part-of-speech and morphological tagging, as well as dependency parsing within the Universal Dependencies framework.[5] The dependency tree generated by UDPipe for the English version of the sentence from Example 1 is illustrated in Figure 2.

In the parsed document, DReUD subsequently searches (i) for explicit intra-sentential discourse relations (separately in coordinating and subordinating structures), and (ii) for explicit inter-sentential discourse relations, using the tree structure, part of speech and morphological information and dependency functions. In all these cases, the system first tries to identify Argument 2 of a relation by searching for a finite verb[6] with properly attached connective tokens. If Argument 1 (again, represented by a finite verb) is also found at the correct position in the tree (possibly with additional connective nodes[7]), the potential connective is compiled from the connective tokens and searched for in the lexicon of discourse connectives.[8] If the connective is found in the lexicon, the discourse relation is marked in the data and it is assigned—for now as a baseline—the most frequent discourse type found for the connective in the lexicon.[9]

A lexicon of discourse connectives is an essential part of the parser, used both for finding explicit relations and for their classification. For Czech, the parser uses the Lexicon of Czech discourse connectives CzeDLex (Mírovský et al., 2017). It contains about 200 entries with detailed manual annotation and corpus-based frequencies of discourse types for individual connectives.

For English, the lexicon DiMLex-Eng (Das et al., 2018) is used. In its current version, it contains 142 entries, incl. corpus-based frequencies of various senses of individual connectives. The original Penn senses were transformed to the Prague discourse types.

5. Conclusion

We have presented a proposal for an annotation scheme and data format for capturing shallow discourse relations within the Universal Dependencies framework, as a part of an effort to bring discourse relations analysis to simple and rapid applicability both in theoretical research and NLP applications.

As a principal part of the effort, the DReUD application is being developed, with its current development state being a work-in-progress baseline model for shallow discourse parsing in Czech and English.

[4] https://lindat.cz/services/udpipe/

[5] UDPipe models have been trained on nearly all UD treebanks, including Czech and English, achieving state-of-the-art performance.

[6] This is a simplified wording for possibly more complex situations, such as a coordination of finite verbs or verb constructions with finite auxiliaries. Also note that the restriction to arguments with finite verbs may be relaxed in future versions.

[7] for complex connectives such as *if ~ then*

[8] For inter-sentential relations, as the first approximation to be improved later, the whole previous tree (i.e., the previous sentence) is chosen as Argument 1.

[9] Such a baseline is actually quite strong and hard to beat without using large language models, as was shown, e.g., in Mírovský and Poláková (2021), or more recently in Pribytkova (2025).

The modularity of the DReUD system and the uniform generality of Universal Dependencies allows for simple addition of further languages. To extend the system to a new language, a respective lexicon of discourse connectives along with at least the most common discourse sense for each connective is needed, or an annotated corpus from which such information could be automatically extracted. At the very least, given the existing resources and language skills of the development team, we plan to add support for Spanish and German.

DReUD is available as a command-line program, web service and a REST API server.[10] Its source code is downloadable under the Mozilla Public License 2.[11]

Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project 26-20719S). 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).

6. Bibliographical References

Massinissa Atmani and Mathieu Lafourcade. 2021. Universal dependencies for multilingual open information extraction. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, pages 24–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of english discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2017. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722.

Hiroshi Kanayama and Ran Iwamoto. 2020. How universal are universal dependencies? exploiting syntax for multilingual clause-level sentiment detection. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4063–4073.

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

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

Jiří Mírovský and Lucie Poláková. 2021. Sense prediction for explicit discourse relations with BERT. In *Proceedings of Sixth International Congress on Information and Communication Technology (ICICT)*, volume 216 of *Lecture Notes in Networks and Systems*, pages 835–842, Singapore. International Congress and Excellence Awards, Springer.

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.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment analysis in twitter with lightweight discourse analysis. In *Proceedings of COLING 2012*, pages 1847–1864.

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

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872.

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

Olga Pribytkova. 2025. Discourse relations of the Prague Discourse Treebank in Universal Dependencies.

[10] https://quest.ms.mff.cuni.cz/dreud/
[11] https://github.com/ufal/dreud/

Dongqi Pu, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. *arXiv preprint arXiv:2305.16784*.

Petr Sgall. 1967. Functional sentence perspective in a generative description. *Prague studies in mathematical linguistics*, 2(1967):203–225.

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

Karin Sim Smith and Lucia Specia. 2018. Assessing crosslingual discourse relations in machine translation. *arXiv preprint arXiv:1810.03148*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the Prague Discourse Treebank 3.0. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1270–1279, Torino, Italy. European Language Resources Association.

Suzan Verberne, LWJ Boves, NHJ Oostdijk, and PAJM Coppen. 2007. Discourse-based answering of why-questions.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

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

7. Language Resource References

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

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

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