Cost-Effective Discourse Annotation

in the Prague Czech–English Dependency Treebank

Jiří Mírovský, Pavlína Synková, Lucie Poláková and **Marie Paclíková**

Charles University, Institute of Formal and Applied Linguistics, Prague, Czech Republic

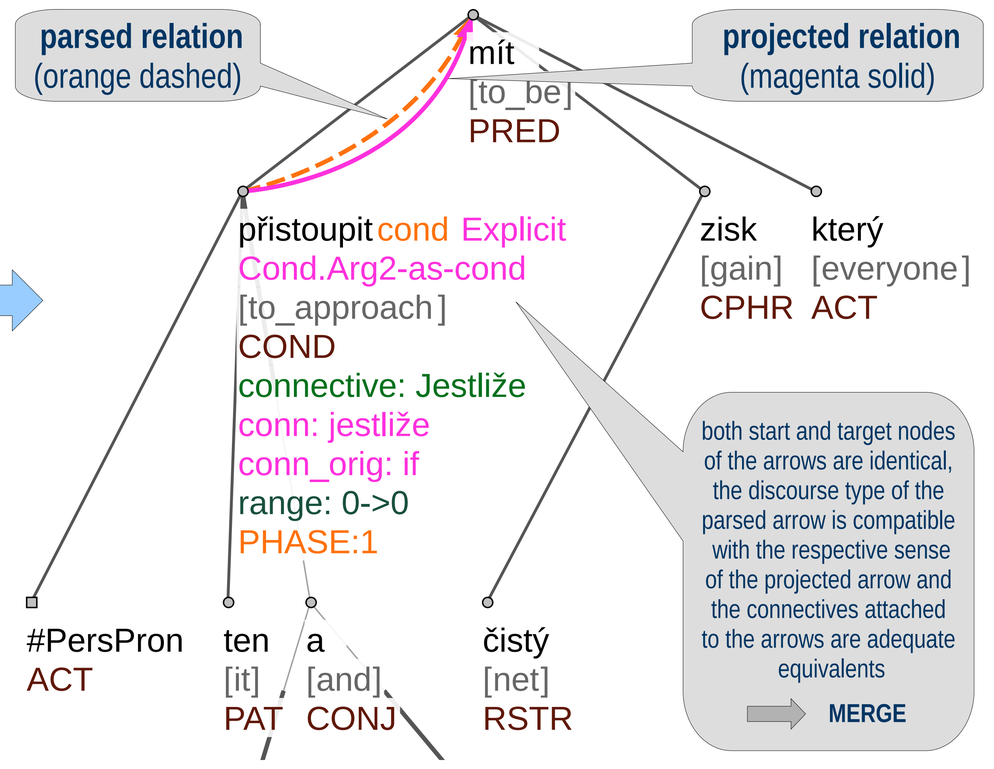
We combine **three resources** to obtain high-quality **discourse annotation** in the Czech part of the PCEDT: (i) **annotation projection** from the Penn Discourse Treebank 3.0, (ii) manual **tectogrammatical** (deep syntax) **representation** of sentences of the corpus, and (iii) the **Lexicon of Czech Discourse Connectives CzeDLex**.

Result: F1 on existence: 0.87, accuracy on types: 78% with Cohen's Kappa 0.73. For comparison, IAA in PDiT 1.0 was: F1: 0.84, acc: 74%, kappa: 0.68.

The Method

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**
7. solving **ambiguous transformation to Penn senses**

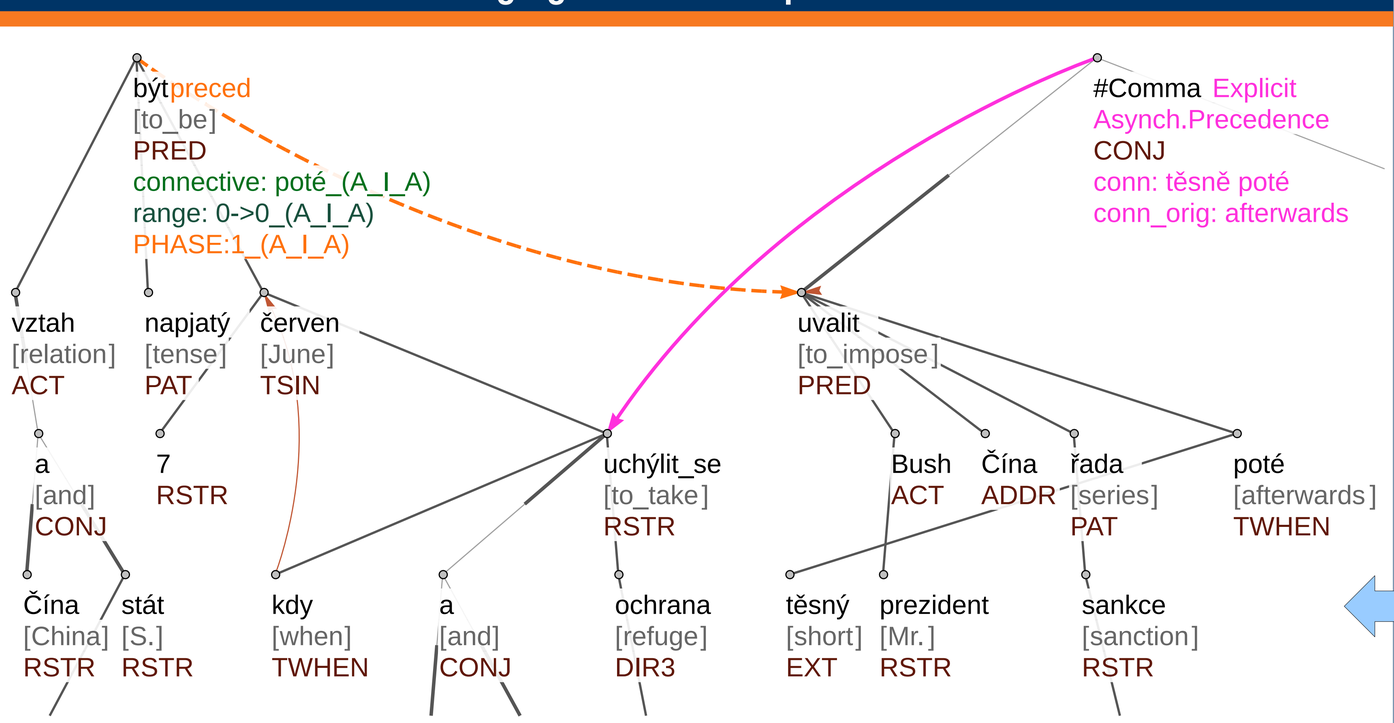
Merging: A Complete Overlap



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*.]

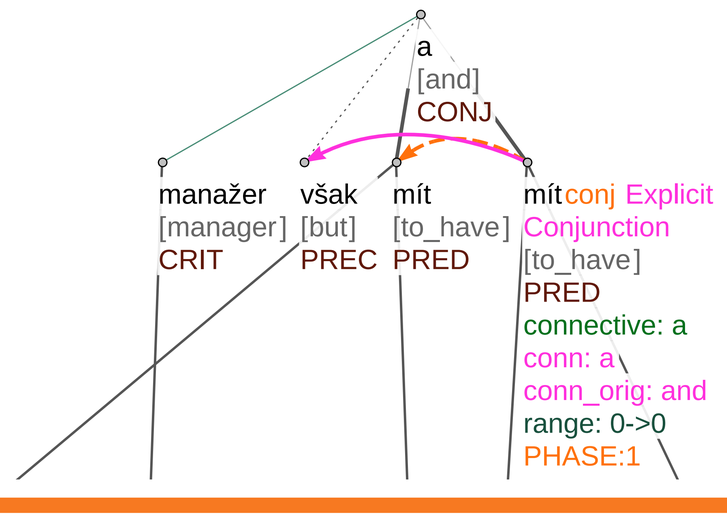
Merging: A More Complex Case



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**.]

Merging: Different Target Nodes



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**.]

See Also

CzeDLex (Lexicon of Czech Discourse Connectives)
https://ufal.mff.cuni.cz/czedlex1.0/

TrEd (Tree Editor)
https://ufal.mff.cuni.cz/tred/

Discourse annotation of the **Prague Czech–English Dependency Treebank** (the Czech part)

49,208 sentences, 28,804 explicit discourse relations
manual intervention:
- 2 thousand positions during various stages of preparation,
- 6 thousand positions checked to verify the existence of a relation and its arguments,
- 8 thousand positions inspected to check the discourse type or sense

Available by the end of 2024 **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; Creative Commons License (CC BY-NC-SA 4.0).

The authors gratefully acknowledge support from the Grantl Agency of the Czech Republic (project "Methods for Rapid Discourse Annotation in Selected Corpora", project number 22-03269S).
This work has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project number LM2023062).