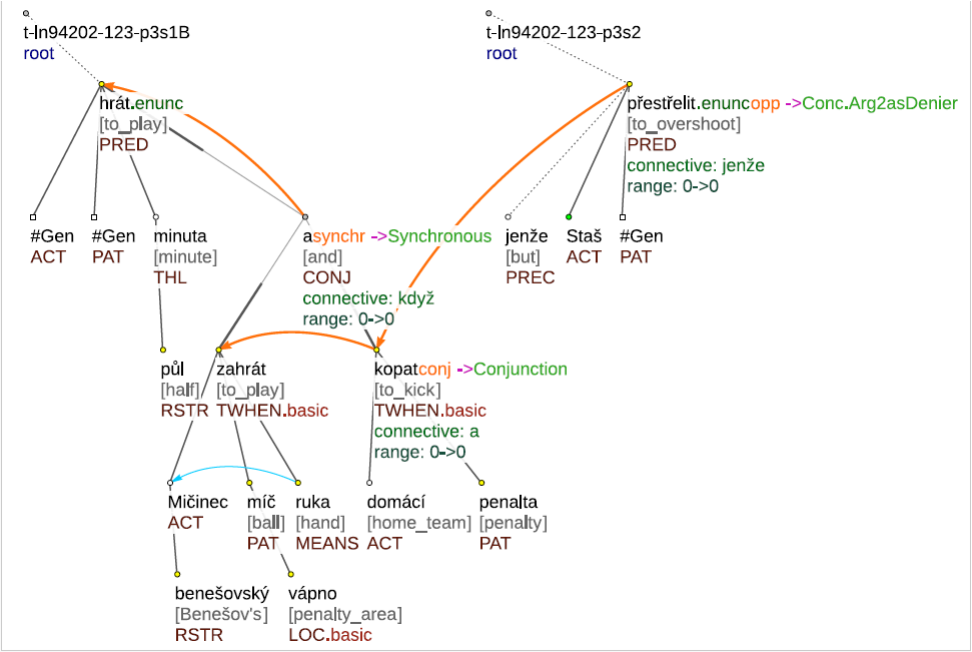
Announcing the Prague Discourse Treebank 3.0

Pavlína Synková, Jiří Mírovský, Lucie Poláková and Magdaléna Rysová

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

- layer of discourse relations annotated on the top of dependency trees of Prague Dependency Treebank
  - approx. 49 000 sentences
  - approx. 22 000 discourse relations



Hrálo se půl minuty, když benešovský Mičinec zahrál míč ve vápně rukou a domácí kopali penaltu.
Jenže Staš přestřelil.

The game was played for half a minute when Benešov's Mičinec played the ball in the penalty area with his hand and the home team kicked a penalty.
But Staš overshot.

PDiT

- discourse relations
  - shallow approach
  - only explicit (marked by primary or secondary connectives)
  - argument formed by at least one clause containg finite verb
  - taxonomy follows Czech syntactic tradition and is inspired by PDTB2.0 taxonomy
- + lists, headings, metatexts, genres (one for each document)

- PDiT 1.0 – primary connectives
- PDiT 2.0 - + secondary connectives
- PDiT 3.0
  - annotation revisions
  - transformation into PDTB3.0 format and taxonomy

PDiT3.0

- annotation revisions
  - revision of pragmatic relations and explication
  - correction based on work on Lexicon of Czech Discourse Connectives
  - annotators´comments
- transformation into PDTB3.0 format and taxonomy
  - discussed in detail in Mírovský et al. (2023), https://ufal.mff.cuni.cz/pbml/120/art-mirovsky-et-al.pdf

PDiT 3.0 – revisions: pragmatic relations

(i) relations that involve some pragmatic phenomenon like subjectivity, complex inferencing, presuppositions etc.

(ii) relations where the form and the meaning do not correspond (but at the same time the relation cannot be interpreted as another semantic relation), including stylistically inappropriate contexts

30 % of pragmatic relations re-annotated, they were annotated as pragmatic erroneously due to lack of contextual knowledge

PDiT 3.0 – revisions: explication

- connectives *totiž*, *vždyť*, *přece* [you see, actually, you know], often implicited in English
- explanation, that is not necessarily given by means of a causal connection, but by means of a more elaborate reformulation of the left argument

52 % of 279 explications were re-annotated as reason-result

PDiT 3.0 – revisions: corrections based on CzeDLex

- CzeDLex – Lexicon of Czech Discourse Connectives
  - built in 2016-2021
  - 200 entries with numerous complex forms and modifications
  - automatically extracted from PDiT2.0, completely manually checked, available online
  - https://ufal.mff.cuni.cz/czedlex/

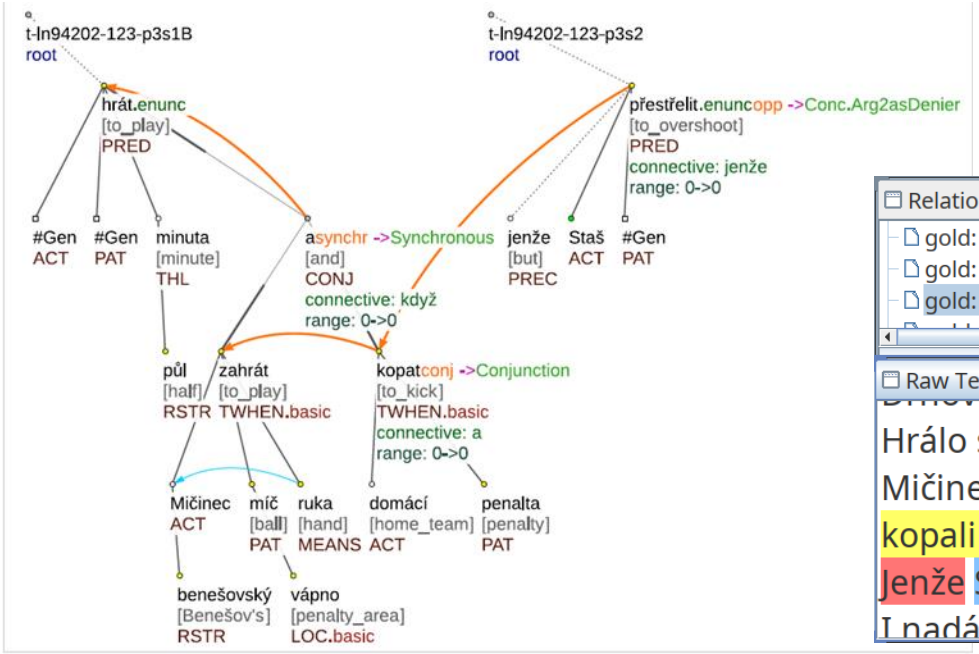
- expressions in connective use
  - false non-connective usages
  - 1,400 instances of 20 expressions revised, 360 new relations,180 other modifications

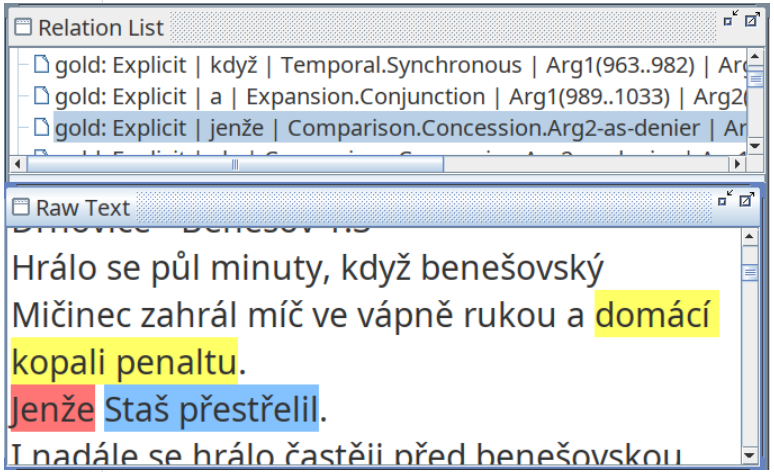
- individual contexts
  - connectives, types, arguments, existence of relations in given context
  - 300 contexts, 140 new relations, 190 other modifications

PDiT 3.0 – revisions: annotators´comments

- 1,100 comments in PDiT2.0 – 240 of them relevant for PDiT3.0
  - mismatch between manual and automatic part of annotation
  - uncertainty about the type of relation or its existence in a given context
  - 40 new relations, 130 other modifications

PDiT 3.0: transformation into PDTB3.0 taxonomy





The game was played for half a minute <u>when</u> Benešov's Mičinec played the ball in the penalty area with his hand *<u>and</u> the home team kicked a penalty.*
<u>But</u> Staš overshot.

PDiT 3.0: transformation into PDTB3.0 taxonomy

PDiT 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

- one-to-one correspondence: automatic transformation
- more senses - one type: loss of information, automatic transformation

  (both types together approx. 42 % of relations)
- one type - more senses: studied and transformed

  a) automatically using linguistic features (approx. 56 % of relations)

  b) manually (pragmatic relations and explication, 2 % of relations)

Summary

- Prague Discourse Treebank 3.0 has
  21,662 discourse relations in 49,428 sentences
- manual annotation, explicit discourse relations
- differences from the previous version:
  - annotation revisions (approx. 4500 context revised, approx. 550 new relations)
  - transformation into PDTB3.0 format and taxonomy

https://ufal.mff.cuni.cz/pdit3.0