Towards a Discourse Corpus of Czech

Lucie Mladová
Šárka Zikánová, Zuzanna Bedřichová,
Jiří Mírovský, Eva Hajičová

Charles University in Prague
Institute of Formal and Applied Linguistics

Corpus Linguistics, Liverpool, July 22, 2009

Outline

- Groundwork in the Area of Discourse Structure
- Methods for the Prague Discourse Annotation
- Present State of the Art

From Syntactic Structure to Discourse Structure

- Condition relation in the syntactic point of view (within a sentence):
  - *If you don't like your meal, don't eat it.*
  - *I will cook pancakes, if you buy eggs.*
  - *If he had married, he wouldn't be such a killjoy now.*
- Condition relation across the sentence boundary:
  - *You don't like your meal? Then don't eat it!!*
  - *I will cook pancakes. But you must buy eggs first.*
  - *He should have married. He wouln't be such a killjoy now.*
- Pragmatic meanings: cause and "pragmatic cause"
  - *John is home because he is sick.*
  - *John is home because the lights are on in the house.*

The Goal: Capturing Discourse Structure

- Connecting of discourse units (clauses, sentences) into a coherent, meaningful sequence of information

How:

- Marking anaphoric chains (textual coreference) and brigding relations
- Marking semantically related text spans - connective properties of certain discourse markers
- Other discourse-relevant information: style of the text, attribution to the speaker/writer, salience, topic-focus articulation, etc.

The Goal: Capturing Discourse Structure

- Connecting of discourse units (clauses, sentences) into a coherent, meaningful sequence of information

How:

- Marking anaphoric chains and brigding relations (textual coreference)
- **Marking semantically related text spans - connective properties of certain discourse markers**
- Other discourse-relevant information: style of the text, attribution to the speaker/writer, salience, topic-focus articulation, etc.

The Goal: Capturing Discourse Structure

Why:

- Large-scale database for:
  - linguistic research grounded on language corpora (the first "discourse corpus" for Czech!)
- NLP applications such as:
  - Automatic text summarization
  - Information retrieval
  - Dialog systems
  - Question answering tasks
  - Automatic annotation of further data

Semantics in Discourse

- Inter-sentential semantic relations (between separate trees) marked by discourse connectives (DCs)
  - *She never stayed at home in the evenings. She often went out with friends, for example.* Instantiation
  - *The method is up to you. Just do it your own way.* Equivalence
- A difficult task: relations represented by implicit connectives (not present on the surface)
  - *He takes bends with great care. He always slows down.* Specification

Resources for Prague Discourse Annotation

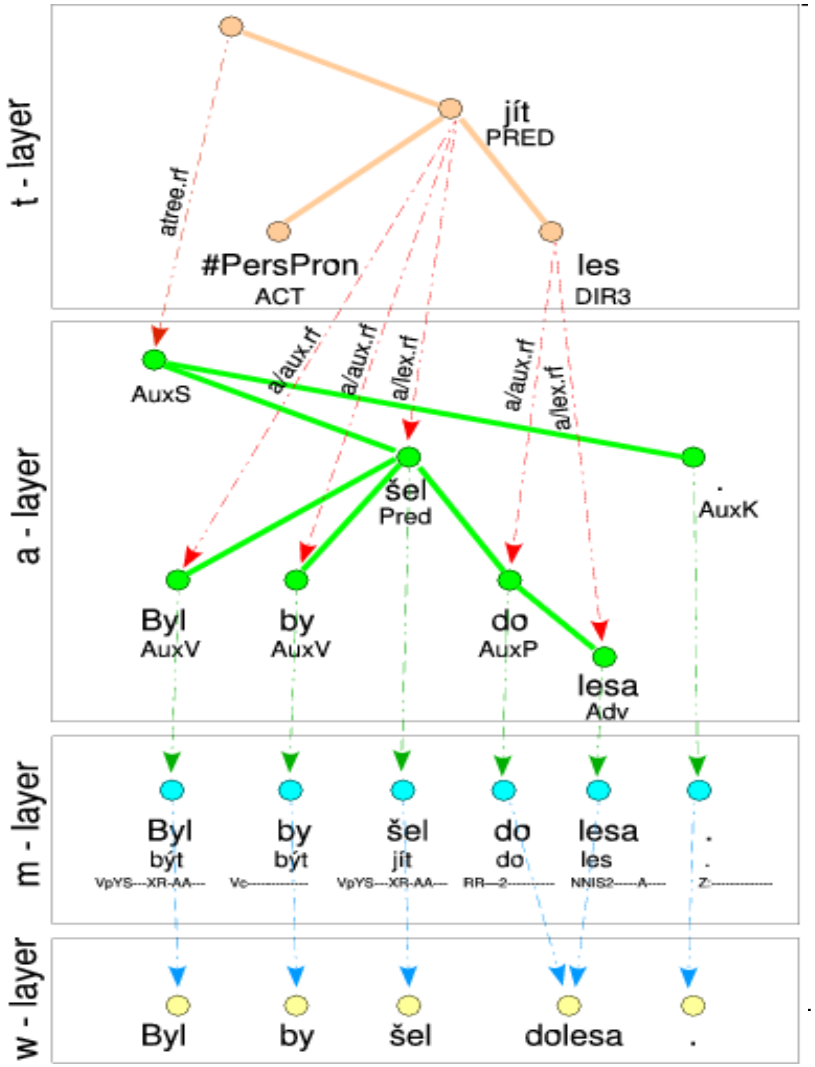
- Present-day syntactico-semantic (i.e. tectogrammatic) annotation in Prague Dependency Treebank 2.0 (PDT) for Czech

- Penn Discourse Treebank 2.0 (PDTB) approach – identifying discourse connectives and their arguments for English

  (Joshi et al. 2008)

→ Prague set of discourse relations inspired by both Prague (syntactico-semantic) and Penn (lexical) approach

Existing Annotation Layers in the Prague Dependency Treebank (PDT 2.0)



4 layers (3 annotation layers + plain text (w-layer))

T-layer (deep syntax or tectogrammar) contains some discourse relations already:

1. coordinations
2. subordinate (dependent) clauses
3. no inter-sentential marking

BUT

some of the discourse connectives are marked by the label PREC (reference to PREceding Context)

Practical Decisions for the Annotation

- Automatically:
  - Extraction of relevant discourse information from the present PDT 2.0, event. automatic preannotation
- Manually:
  - Subset of texts with rich connective property
  - Only explicit DCs in the first step
  - Only those with clausal arguments
  - Mainly PREC-labeled, but checking for other possible DCs

Discourse Sense Tags

- TEMPORAL
  - precedence-succession
  - simultaneity
- CONTRAST
  - confrontation
  - opposition
  - restrictive opposition
  - concession
  - correction
  - gradation
- CONTINGENCY
  - reason-result
  - textual condition
  - purpose
  - explication
- EXPANSION
  - conjunction
  - instantiation
  - specification
  - equivalence
  - generalization
  - conjunctive alternative
  - disjunctive alternative

Annotation example

I will cook pancakes. *But you must buy eggs first.*

Annotation example

I will cook pancakes. ***But*** *you must buy eggs first.*

Annotation example

I will cook pancakes. ***But*** *you must buy eggs first.*

Annotation example

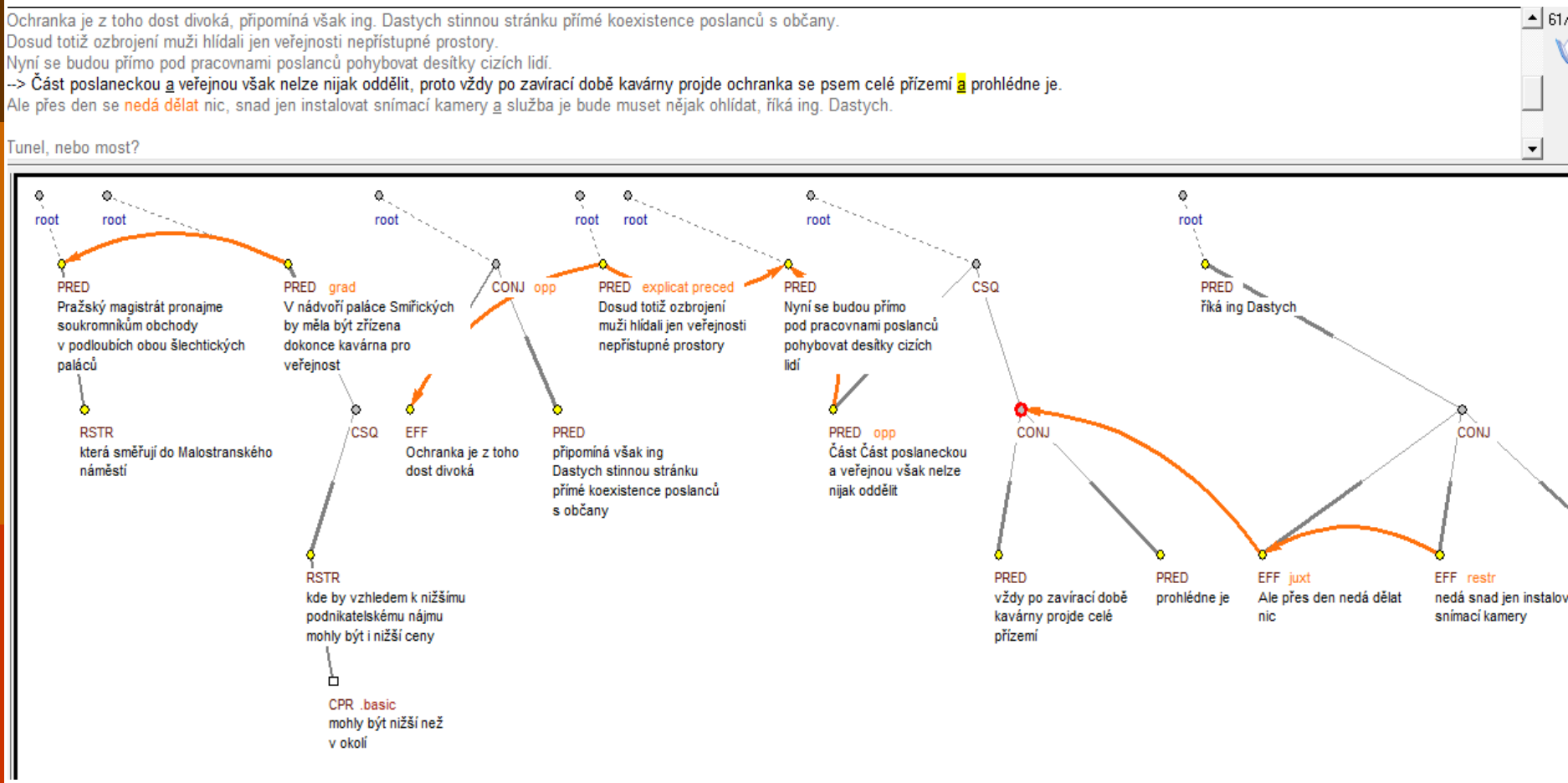
I will cook pancakes. ***But*** *you must buy eggs first.*

condition

He takes bends with great care. *He always slows down.*

specification

Annotation on the Trees (TrEd Tool)



Present State of the Art

- Methodology and tool preparations completed
- Evaluate preliminary annotations for IAA (3 annotators) in:
  1. identifying discourse connectives
  2. identifying their arguments
  3. assigning discourse-semantic label to the relation
- Initiate full manual annotations (demo version app. 5000 Czech discourse-annotated sentences in spring 2010)
- Verify the linguistic decisions made on a bigger amount of annotated data
- 2011/2012: The PDT 2.0 (app. 50 000 sentences) with a new annotation layer (discourse structure, coreference)

Thank you for your attention!

mladova@ufal.mff.cuni.cz

The work on this project is supported by the grants GA201/09/H057 and GA405/09/0729 of the Grant Agency of the Czech Republic, and GAUK 103609 of the Grant Agency of the Charles University Prague