From Sentence to Discourse

Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank

Lucie Mladová, Šárka Zikánová, Eva Hajičová

Institute of Formal and Applied Linguistics
Charles University in Prague

May 28, 2008

Outline

1. Language Resources and Theoretical Background
    - Outline
    - Prague Dependency Treebank
    - Penn Discourse TreeBank
2. Building a Discourse Corpus
    - General Principles
    - Specific Issues
3. Conclusion
    - Current and Future Work

Prague Dependency Treebank

- A corpus of Czech journalistic texts (approx. 2 million word units)
- The annotation scheme: from structure to function - 3 layers of annotation:
  - Morphological layer
  - Analytical layer (surface syntax)
  - Tectogrammatical layer (deep syntax and semantics)

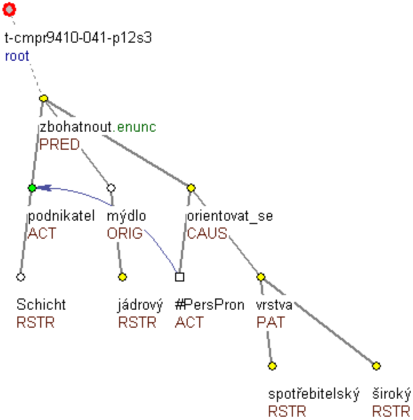
The tectogrammatical representation

Sentence structure - dependency trees
Syntactico-semantic labels - functors
Topic-focus articulation

Coreference

Tectogrammatical Tree Structure

An example of a tectogrammatical tree (a single-sentence representation)

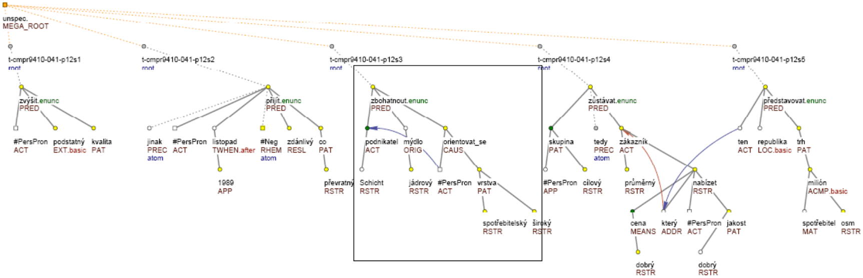


”Podnikatel Schicht zbohatl na jádrovém mýdle, protože se orientoval na nejširší spotřebitelskou vrstvu.”

”The entrepreneur Schicht got rich on grain soap because he concentrated on the widest consumer rank.”

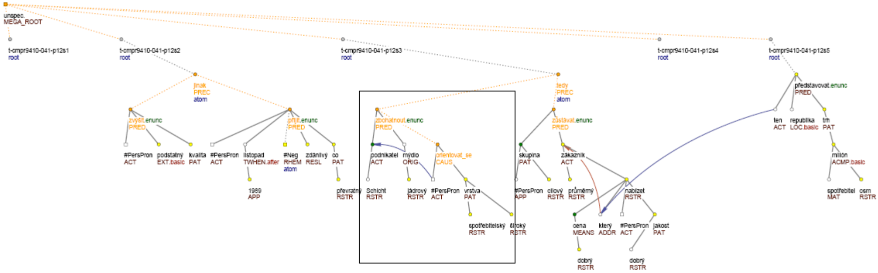
The Idea of a Discourse Treebank

A proposal of a megatree (a five-sentence-discourse representation)



The Idea of a Discourse Treebank

A proposal of a megatree (a five-sentence-discourse representation)



Penn Discourse TreeBank

For Comparison:

- Discourse annotation of WSJ texts (version 2.0 of PDTB released 2008)
- Structuring of the texts by lexical items - discourse connectives

Discourse annotation in Penn

Description of the **discourse connectives** and their **arguments**

Each discourse connective takes exactly two arguments

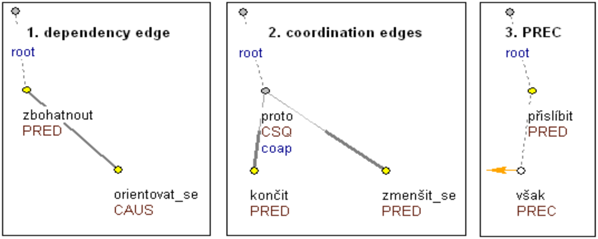
Semantic classification of discourse relations - set of semantic labels

From Tectogrammatics to Discourse

- Prague underlying syntax annotation - some discourse relations already captured
- Some of Prague tectogrammatical functors - discourse semantics
- Discourse annotations only a part of the new layer of PDT 3.0, also included:
    - Topic-focus articulation (TFA)
    - Named entities
    - Extended coreference annotations
    - Other textual relations
- Megatree representation - update of the current tool TrEd (Tree Editor)
- No "lower" information lost

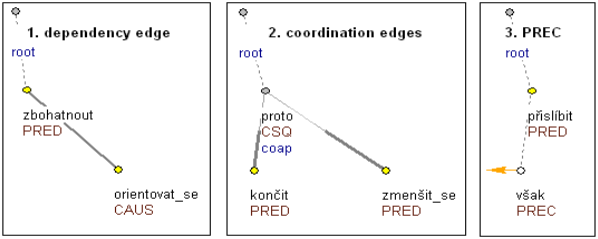
Three Types of Capturing a Possible Discourse Relation

in Prague Dependency Treebank



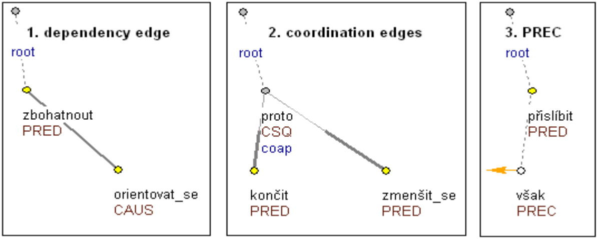
1. **Dependency** (tectogrammatical functors for verb free modifiers such as: CAUS, COND, AIM, CNCS, TWHEN, LOC, DIR, MANN, ACMP, REG etc.) **but not** for inner participants of the valency frame of the verb (ACT, PAT, ADDR, ORIG, EFF)

Three Types of Capturing a Possible Discourse Relation
in Prague Dependency Treebank



1. **Dependency** (tectogrammatical functors for verb free modifiers such as: CAUS, COND, AIM, CNCS, TWHEN, LOC, DIR, MANN, ACMP, REG etc.) **but not** for inner participants of the valency frame of the verb (ACT, PAT, ADDR, ORIG, EFF)
2. **Coordination** (functors CONJ, GRAD, DISJ, ADVS, CSQ, CONFR, OPER, REAS, APPS etc.), **but not** coordination of minor units (John and Mary)!

Three Types of Capturing a Possible Discourse Relation
in Prague Dependency Treebank



1. **Dependency** (tectogrammatical functors for verb free modifiers such as: CAUS, COND, AIM, CNCS, TWHEN, LOC, DIR, MANN, ACMP, REG etc.) **but not** for inner participants of the valency frame of the verb (ACT, PAT, ADDR, ORIG, EFF)
2. **Coordination** (functors CONJ, GRAD, DISJ, ADVS, CSQ, CONFR, OPER, REAS, APPS etc.), **but not** coordination of minor units (John and Mary)!
3. The **PREC** functor

PREC - reference to PREceding Context

- An expression marked with PREC indicates a simple presence of a discourse relation:

  Hence PREC, I am happy.

  An isolated research, however PREC, cannot have good results.

- PREC applies primarily to units across the sentence boundaries (is "anaphoric")

PREC - reference to PREceding Context

- An expression marked with PREC indicates a simple presence of a discourse relation:

  Hence PREC, I am happy. CSQ - consequence

  An isolated research, however PREC, cannot have good results. ADVS - adversative

- PREC applies primarily to units across the sentence boundaries (is "anaphoric")
- Needs to be subclassified

Comparison of Penn and Prague Semantic Labels

- Prague tectogrammatical functors not marked yet explicitly as discourse sense labels
- Penn labels - hierarchical organization, functors non-hierarchical

Comparison of Penn and Prague Semantic Labels

- Prague tectogrammatical functors not marked yet explicitly as discourse sense labels
- Penn labels - hierarchical organization, functors non-hierarchical
  1. *[Jakou povahu jsi měl], než [jsi přišel o práci]?*
     [What had you been like] before [you lost your job]?
     discourse connective = before
     PDTB: temporal - asynchronous - precedence
     PDT: functor TWHEN - temporal, subfunctor BEFORE

Comparison of Penn and Prague Semantic Labels

- Prague tectogrammatical functors not marked yet explicitly as discourse sense labels
- Penn labels - hierarchical organization, functors non-hierarchical

1. *[Jakou povahu jsi měl], než [jsi přišel o práci]?*
   [What had you been like] before [you lost your job]?
   discourse connective = before
   PDTB: temporal - asynchronous - precedence
   PDT: functor TWHEN - temporal, subfunctor BEFORE

2. *[Bud' půjdeme do kina], nebo [zůstaneme doma].*
   [Either we'll go to the cinema], or [we'll stay at home].
   discourse connective = or (disjunctive meaning)
   PDTB: expansion - alternative - disjunctive
   PDT: functor DISJ - disjunctive

Comparison of Penn and Prague Semantic Labels

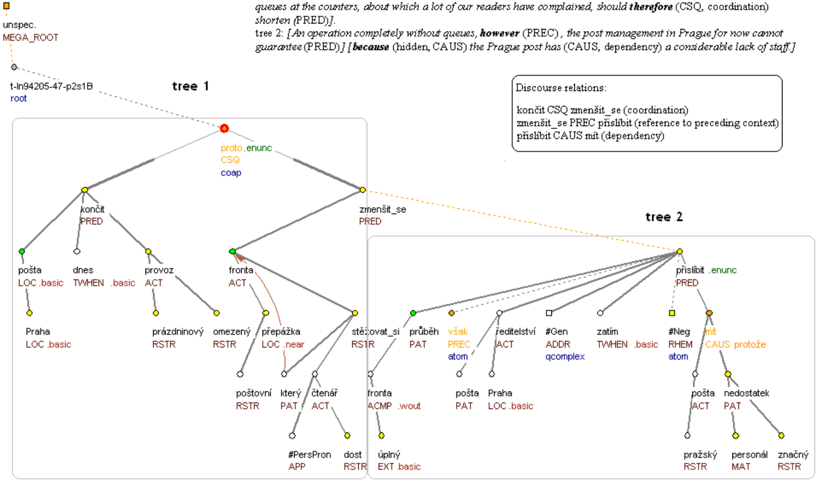
- Prague tectogrammatical functors not marked yet explicitly as discourse sense labels
- Penn labels - hierarchical organization, functors non-hierarchical

  1. *[Jakou povahu jsi měl], než [jsi přišel o práci]?*
     [What had you been like] before [you lost your job]?
     discourse connective = before
     PDTB: temporal - asynchronous - precedence
     PDT: functor TWHEN - temporal, subfunctor BEFORE

  2. *[Bud' půjdeme do kina], nebo [zůstaneme doma].*
     [Either we'll go to the cinema], or [we'll stay at home].
     discourse connective = or (disjunctive meaning)
     PDTB: expansion - alternative - disjunctive
     PDT: functor DISJ - disjunctive

  3. *[...] A [potom odešel].*
     [...] And [then he left].
     discourse connective = and
     PDTB: expansion - conjunction
     PDT: functor PREC (no discourse semantics marked)

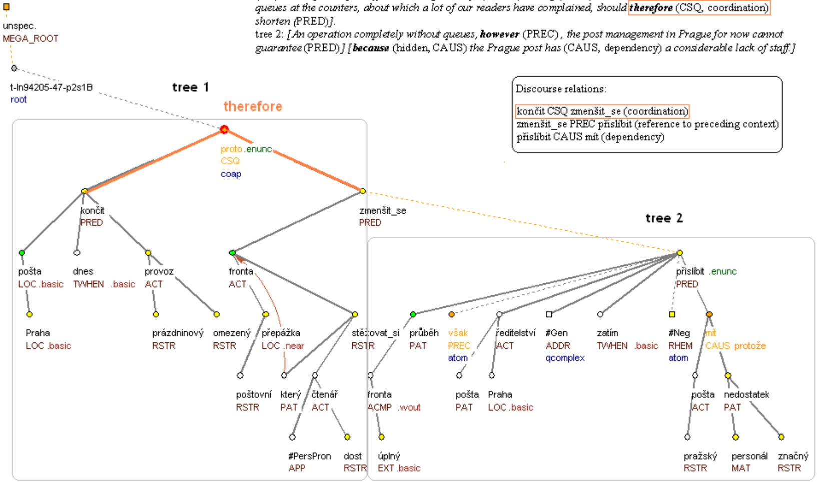
(Lit.) tree 1: *[At the post offices in Prague today,* (there is) *ending* (PRED) *the restricted holiday operation], [the queues at the counters, about which a lot of our readers have complained, should* ***therefore*** (CSQ, coordination) *shorten* (PRED)*].*
tree 2: *[An operation completely without queues,* ***however*** (PREC) *, the post management in Prague for now cannot guarantee* (PRED)*] [****because*** (hidden, CAUS) *the Prague post has* (CAUS, dependency) *a considerable lack of staff.]*



Discourse relations:

končit CSQ zmenšit_se (coordination)
zmenšit_se PREC přislíbit (reference to preceding context)
přislíbit CAUS mít (dependency)

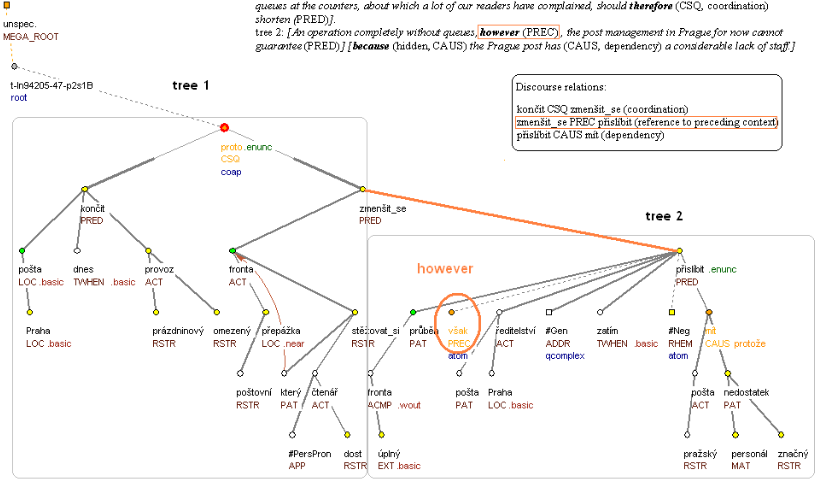
(Lit.) tree 1: *[At the post offices in Prague today,* (there is) *ending* (PRED) *the restricted holiday operation], [the queues at the counters, about which a lot of our readers have complained, should* ***therefore*** (CSQ, coordination) *shorten* (PRED)*].*
tree 2: *[An operation completely without queues,* ***however*** (PREC) *, the post management in Prague for now cannot guarantee* (PRED)*] [****because*** (hidden, CAUS) *the Prague post has* (CAUS, dependency) *a considerable lack of staff.]*



Discourse relations:

končit CSQ zmenšit_se (coordination)
zmenšit_se PREC přislíbit (reference to preceding context)
přislíbit CAUS mít (dependency)

(Lit.) tree 1: *[At the post offices in Prague today,* (there is) *ending* (PRED) *the restricted holiday operation], [the queues at the counters, about which a lot of our readers have complained, should* ***therefore*** (CSQ, coordination) *shorten* (PRED)*].*
tree 2: *[An operation completely without queues,* ***however*** (PREC)*, the post management in Prague for now cannot guarantee* (PRED)*] [****because*** (hidden, CAUS) *the Prague post has* (CAUS, dependency) *a considerable lack of staff.]*



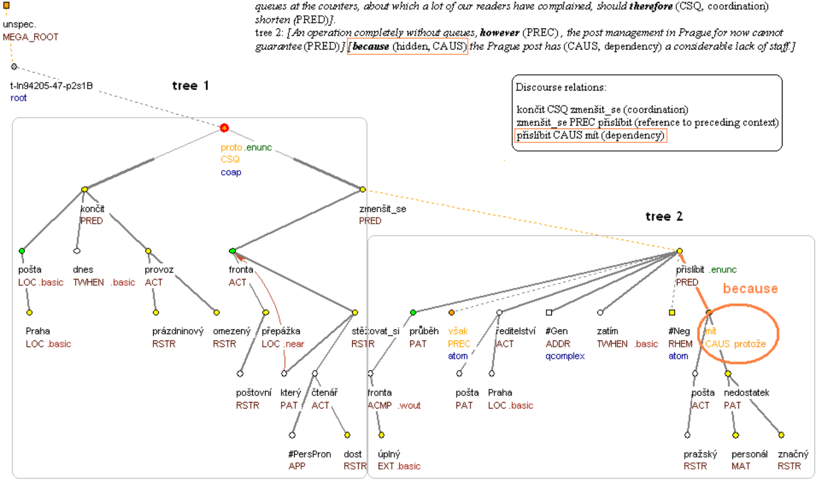
Discourse relations:

končit CSQ zmenšit_se (coordination)
zmenšit_se PREC přislíbit (reference to preceding context)
přislíbit CAUS mít (dependency)

(Lit.) tree 1: *[At the post offices in Prague today,* (there is) *ending* (PRED) *the restricted holiday operation], [the queues at the counters, about which a lot of our readers have complained, should* ***therefore*** (CSQ, coordination) *shorten* (PRED)*].*
tree 2: *[An operation completely without queues,* ***however*** (PREC) *, the post management in Prague for now cannot guarantee* (PRED) *] [****because*** (hidden, CAUS) *the Prague post has* (CAUS, dependency) *a considerable lack of staff.]*

Discourse relations:

končit CSQ zmenšit_se (coordination)
zmenšit_se PREC přislíbit (reference to preceding context)
přislíbit CAUS mít (dependency)



Open Questions

- Delimitation of the discourse units
  - Parcelling
  - Verbless clauses
  - Parentheses
  - Nominalizations
- Binarity of the discourse connectives (as in PDTB)
- Language-specific discourse phenomena
- Etc.

Current Issues Worked on

- Lists of English and Czech expressions with the possible PREC function
- Comparison of PDTB 2.0 sense label set with the Prague functors
- Creating of the megatree context for tree adjoining experiments, mapping both linguistic and technical conditions
- Experimental annotations of the PDT data (Czech) and NAP-Corpus dialog data (English)

Future Work

- Revision and extension/reduction of the functors with respect to the Penn sense label set
- Work with both written (PDT, WSJ) and spoken (dialog, NAP) texts
- Work with both Czech and English data
- Build on the previous linguistic work (tree structures, underlying syntax, coreference and TFA annotations)

  ⟶ **Building a consistent annotation scenario for discourse**

Acknowledgements

Thank you for your attention!

{mladova,zikanova,hajicova}@ufal.mff.cuni.cz

The work on this project is supported by the grants CKL (LC536), EU Companions (207-55/6694) and MSM-0021620838.