Prague Dependency Treebank: Introduction – trees, dependency

Markéta Lopatková, Jiří Mírovský

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz
NPFL075 Prague Dependency Treebank

Lectures:
Markéta Lopatková

Practical sessions:
Jiří Mírovský

http://ufal.mff.cuni.cz/course,npfl075

Requirements:
• Homework (40%)
• Activity (10%)
• Final test (50%)

Assessment:
• excellent (= 1) ≥ 90%
• very good (= 2) ≥ 70%
• good (= 3) ≥ 50%
Prague Dependency Treebank

Collection of:
• linguistically annotated data (Czech)
• tools and data format(s)
• documentation

Another point of view:
• annotation scheme
• framework for annotation of different languages
• underlying linguistic theory (Functional Generative Description)
Prague Dependency Treebank

Collection of:
- linguistically annotated data (Czech)
- tools and data format(s)
- documentation

Another point of view:
- annotation scheme
- framework for annotation of different languages
- underlying linguistic theory (Functional Generative Description)

What about other/similar approaches:
- HamleDT
- Universal Dependencies
Outline of the lecture

- trees (graph theory and data format)
- phrase structure trees and dependency trees
- dependency and non-dependency relations
- non-projectivity
How to capture sentence structure?

Gate *receipts are* only the Cowboys' second largest source of cash.
Graph theory: tree

tree (graph theory):
definition:
- finite graph \( \langle N, E \rangle \), \( N \sim \) nodes/vertices, \( E \sim \) edges \{n_1,n_2\}
- connected
- no cycles, no loops
- no more than 1 edge between any two different nodes

⇔ (undirected) graph
any two nodes are connected by exactly one simple path
**Graph theory: tree**

**tree** (graph theory):

definition:

- finite graph \( \langle N, E \rangle \), \( N \sim \) nodes/vertices, \( E \sim \) edges \( \{n_1,n_2\} \)
- connected
- no cycles, no loops
- no more than 1 edge between any two different nodes

⇔ (undirected) graph

any two nodes are connected by exactly one simple path

**rooted tree**

- rooted ⇒ orientation (i.e., edges ordered pairs \([n_1,n_2]\))
Graph theory: tree

tree (graph theory):

definition:
- finite graph \( \langle N, E \rangle \), \( N \sim \) nodes/vertices, \( E \sim \) edges \{n_1,n_2\}
- connected
- no cycles, no loops
- no more than 1 edge between any two different nodes

\( \iff \) (undirected) graph
- any two nodes are connected by exactly one simple path

rooted tree
- rooted \( \Rightarrow \) orientation (i.e., edges ordered pairs \([n_1,n_2]\))

directed tree ... directed graph
- which would be tree
  - if the directions on the edges were ignored, or
  - all edges are directed towards a particular node \( \sim \) the root
Data structure: tree

tree as a data structure:

- rooted tree (as in graph theory)
- all edges are directed from a particular node ~ the root
Data structure: tree

tree as a data structure:

- rooted tree (as in graph theory)
- all edges are directed from a particular node ~ the root
- (linear) ordering of nodes: the children of each node have a specific order
Data structure: tree (properties)

**tree as a data structure:**

- "tree-ordering" $D$ ... partial ordering on nodes
  \[ u \leq v \iff \text{the unique path from the root to } v \text{ passes through } u \]
  (weak ordering ~ reflexive, antisymmetric, transitive)

- "linear ordering" ... (partial) ordering on nodes
  (strong ordering ~ antireflexive, asymmetric, transitive)
Tree-based structures in CL

two types of tree-based structures in CL:

- phrase structure tree / constituent structure tree
- dependency tree
My brother often sleeps in his study.

Phrase structure tree (definition)

\[ T = \langle N, D, Q, P, L \rangle \]

\langle N, D \rangle \ldots \textit{rooted tree, directed}

\( Q \) \ldots lexical and grammatical categories

\( L \) \ldots labeling function \( N \rightarrow Q \)

\( D \) \ldots oriented edges (branches)

\( \sim \) relation on lex. and gram. categories

\textit{dominance relation}

\textit{precedence}

\textit{partial strong linear ordering}
Phrase structure tree (definition)

\[ T = \langle N, D, Q, P, L \rangle \]

\[ \langle N, D \rangle \ldots \textit{rooted tree, directed} \]

\[ Q \ldots \text{lexical and grammatical categories} \]

\[ L \ldots \text{labeling function } N \rightarrow Q \]

\[ D \ldots \text{oriented edges (branches)} \]

\[ \sim \text{ relation on lex. and gram. categories} \]

\textit{dominance relation}

\[ P \ldots \text{relation on } N \sim \text{ (partial strong linear ordering)} \]

\textit{relation of precedence}

Relating dominance and precedence relations:

- \textit{exclusivity} condition for D and P relations
- ‘nontangling’ condition
Phrase structure tree (relation P)

- **exclusivity** condition for D and P relations
  \[ \forall x,y \in N \text{ holds: } ( [x,y] \in P \lor [y,x] \in P ) \iff ( [x,y] \not\in D \land [y,x] \not\in D ) \]

Phrase structure tree (relation P)

```
S
  NP
    Det NP
      my brother
  VP
    Adv VP
      often sleeps in
        Prep NP
          Det NP
            his study
```
Phrase structure tree (relation P)

- **exclusivity** condition for D and P relations
  \[ \forall x,y \in N \text{ holds: } ([x,y] \in P \lor [y,x] \in P) \iff ([x,y] \notin D \land [y,x] \notin D) \]

- **‘nontangling’** condition
  \[ \forall w,x,y,z \in N \text{ holds: } ([w,x] \in P \land [w,y] \in D \land [x,z] \in D) \Rightarrow ([y,z] \in P) \]
Phrase structure tree (relation P)

- **exclusivity** condition for D and P relations
  \[ \forall x,y \in N \text{ holds: } ([x,y] \in P \lor [y,x] \in P) \iff ([x,y] \notin D \land [y,x] \notin D) \]

- **‘nontangling’** condition
  \[ \forall w,x,y,z \in N \text{ holds: } ([w,x] \in P \land [w,y] \in D \land [x,z] \in D) \Rightarrow ([y,z] \in P) \]

\[ T = \langle N,D,Q,P,L \rangle \text{ phrase structure tree} \]
  - \[ \forall x,y \in N \text{ siblings } \Rightarrow [x,y] \in P \]
  - the set of its leaves is totally ordered by P
Phrase structure tree

Pros

- derivation history / ‘closeness’ of a complementation
- coordination, apposition
- CFG-like
- derivation of a grammar
Phrase structure tree

derivation history / ‘closeness’:

... *often sleeps in his study*

... *often sleeps in his study*
Phrase structure tree

**Pros**
- derivation history / ‘closeness’ of a complementation
- coordination, apposition
- CFG-like
- derivation of a grammar

**Contras**
- complexity (number of non-terminal symbols)
- complement (‘two dependencies’)
  - přiběhl bos
  - [(he) arrived barefooted]
- free word order
discontinuous ‘phrases’
non-projectivity
Phrase structure tree

discontinuous ‘phrases’: solution for English

Mary will eat bread.

What will Mary eat?

[Diagram of Phrase Structure Tree]

S

NP

N

Mary

VP

AuxV

V

eat

bread

S

VP

NP

N

what

AuxV

will

Mary

V

eat
discontinuous ‘phrases’: solution for English

Mary will eat bread. What will Mary eat?
discontinuous ‘phrases’: solution for English

Mary will eat bread.

What will Mary eat?
Po babiččině příjezdu půjdou rodiče do divadla.
[After grandma's arrival
the parents will go to the theatre.]
Corpora with phrase structure trees

- Penn Treebank (1995)
  http://www.cis.upenn.edu/~treebank/
  Penn Arabic Treebank, Penn Chinese Treebank
- International English Treebank (ICE)
  http://ice-corpora.net/ice/index.htm
- Paris 7
  http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php
- Szeged Treebank 2.0
- many many others
Dependency tree
My brother often sleeps in his study.

Dependency tree (definition)

\[ T = \langle N, D, Q, WO, L \rangle \]

\( \langle N, D \rangle \) … rooted tree, directed

Q … lexical and grammatical categories

L … labeling function \( N \rightarrow Q \)

D … oriented edges ~ relation on lex. and gram. categories

‘dependency’ relation

WO … relation on \( N \sim \) (strong total ordering on \( N \)) …

word order

\[ \text{sleeps.Pred} \]

\[ \text{brother.Sb} \]

\[ \text{often.Adv} \]

\[ \text{in.AuxP} \]

\[ \text{my.Atr} \]

\[ \text{study.Adv} \]

\[ \text{his.Atr} \]
Dependency tree

Pros
• economical, clear
  (complex labels, ‘word’~ node)
• free word order
• head of a phrase

Contras
• no derivation history / 'closeness'
• coordination, apposition
• complement

PDT – Intro

Lopatková
Dependency tree

discontinuous ‘phrases’: no problem

Mary will eat bread.  
What will Mary eat?

diagram:

```
  eat.Pred
   /   
Mary.Sb  will.AuxV

bread.Obj
```

```
  eat.Pred
   /   
What.Obj  will.AuxV

Mary.Sb
```
Po babiččině příjezdu půjdou rodiče do divadla.
[After grandma's arrival the parents will go to the theatre.]
Corpora with dependency trees

- PropBank (1995)  
  [http://propbank.github.io/](http://propbank.github.io/)

- family of Prague dependency treebanks: Czech, Arabic, English  

- HamleDT project (from 2012)  

- **Universal Dependencies** (from 2013)  
  [http://universaldependencies.org/](http://universaldependencies.org/)

- Danish Dep. Treebank  

- Finnish: Turku Dependency Treebank  
  [http://bionlp.utu.fi/fintreebank.html](http://bionlp.utu.fi/fintreebank.html)

- Negra corpus  

- TIGERCorpus  
  [http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html/](http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html/)

- SynTagRus Dependency Treebank for Russian
Dependency and non-dependency relations
Dependency and non-dependency relations

edges ~ dependency relations (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- long discussion, number of linguistic criteria

E.g., each complete subtree must be a “constituent“, i.e., it must allow for several constructions like topicalization, proform substitution, ….;

Mary will eat bread.

Topicalization:
… and eat Mary certainly will.

Proform substitution:
Mary will do so. (do=eat)

Answer fragment:
What will Mary do? Eat.

VP-ellipsis:
Peter will eat and Mary will, too.

⇒ lexical verb should be a dependent
Dependency and non-dependency relations

edges ~ dependency relations (prototypically)

• dependency relation: binary relation

• governing/modified unit (head) – dependent/modifying unit (modifier)

• PDT criterion: possible reduction
  … dependent member of the pair may be deleted
  while the distributional properties are preserved (→ correctness is preserved)

PDT – Intro Lopatková
Dependency and non-dependency relations

edges ~ dependency relations (prototypically)

• dependency relation: binary relation

• governing/modified unit (head) – dependent/modifying unit (modifier)

• PDT criterion: possible reduction

  … dependent member of the pair may be deleted
  while the distributional properties are preserved (correctness is preserved)

• endocentric constructions … OK

  malý stůl → stůl
  přišel včas → přišel
  (přišel) velmi brzo → (přišel) brzo
  
  small table → table
  he came in time → he came
  (he came) very soon → (he came) soon
Dependency and non-dependency relations

dependency relations (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- PDT criterion: possible reduction
  - dependent member of the pair may be deleted
    - while the distributional properties are preserved (correctness is preserved)
  - endocentric constructions … OK
  - exocentric constructions … principle of analogy on word classes

*Prší.* [(lt) rains.] … ∃ subjectless verbs
⇒ *Král zemřel.* [The king died.] … a verb rather than a noun is the head

*The girl painted a bag. → The girl painted.* … ∃ objectless verbs
⇒ *The girl carried a bag* … an object is considered as depending on a verb
Dependency and non-dependency relations

edges ~ dependency relations (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- PDT criterion: possible reduction
  … dependent member of the pair may be deleted
  while the distributional properties are preserved (→ correctness is preserved)
  - endocentric constructions … OK
  - exocentric constructions … principle of analogy on word classes

PLUS technical considerations

e.g.: prepositions are below nouns;
  auxiliary verbs are (typically) below content verbs
Dependency and non-dependency relations

BUT also other relations:

- coordination … "multiplication" of a single syntactic position
  - different referents
  - coordination of sentence members / sentences
    - My sister Mary and John came late.
    - Mary came in time but John was late.
    - I can't leave since it hasn't stopped raining yet.
    - Nemohu odejít, neboť ještě nepřestalo pršet.

- coordination may be embedded
  - nice and romantic towers and castles
  - krásné a romantické hrady a zámky
Dependency and **non-dependency** relations

BUT also other relations:

- **coordination** … "multiplication" of a single syntactic position
  - different referents
  - coordination of sentence members / sentences
    - *My sister Mary and John came late.*
    - *Mary came in time but John was late.*
    - *I can't leave since it hasn't stopped raining yet.*
    - *Nemohu odejít, neboť ještě nepřestalo pršet.*

- coordination may be embedded
  - *nice and romantic towers and castles*  
  - *krásné a romantické hrady a zámky*

- **apposition** … "multiplication" of a single syntactic position
  - identical referent
    - *Charles IV, Holy Roman Emperor*  
    - *The Hobbit, or There and Back Again*
Dependency and non-dependency relations

BUT also other relations:

- coordination … "multiplication" of a single syntactic position
  - different referents
  - coordination of sentence members / sentences
  - coordination may be embedded

- apposition … "multiplication" of a single syntactic position
  - identical referent

necessary to enrich the data structure
Coordination/apposition in dependency trees

PDT 2.0:

'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- connecting node (= node for coordinating / appositing conjunction)
Coordination/apposition in dependency trees

PDT 2.0:

'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- **connecting node** (= node for coordinating / appositing conjunction)
- **effective parent** (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
Coordination/apposition in dependency trees

PDT 2.0:

'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- connecting node (= node for coordinating / appositing conjunction)
- effective parent (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
- members of a connecting construction (= nodes that are coordinated / are in apposition)
  - is_member
Coordination/apposition in dependency trees

PDT 2.0:

'connecting constructions' ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- **connecting node** (= node for coordinating / appositing conjunction)
- **effective parent** (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
- **members of a connecting construction** (= nodes that are coordinated / are in apposition)
  - **is_member**
- **effective child(ren)** ... modification(s) of the individual member of the connecting construction + common/shared modifier(s)
- ‘pass-through’ nodes
The center will gather and distribute the information on tenders and state commissions in this country as well as in abroad.
Coordination/apposition in dependency trees

PDT 2.0:
- embedded connecting constructions \[\rightarrow\] recursivity

- \textit{TrEd} (Tree Editor, Pajdas):
  functions \texttt{GetEChildren, GetEParents}
Coordination/apposition in dependency

**Universal Dependencies**: version 1 (2014):

the first conjunct

~ the head of all following conjuncts
~ the head of any intervening coordinating conjunctions and punctuation

(Slides stolen from Daniel Zeman)
Coordination/apposition in dependency

*Universal Dependencies*:

version 1 (2014):

the first conjunct

~ the head of all following conjuncts

~ the head of any intervening coordinating conjunctions and punctuation

version 2 (2016):

• the first conjunct ~ the head of all following conjuncts

• attach coordinating conjunctions and punctuation to the immediately succeeding conjunct (instead of the first)

(Slides stolen from Daniel Zeman)
Mel'čuk (1988):

- ‘grouping’ (G) … treating the first conjunct as the head
- problem: shared modification
  vs. modification of a single member

*Hubení ((mladí muži), vojáci a starci)*
[Thin young men, soldiers and old-men]
Coordination/apposition in dependency trees

Petkevič (1995) ... formal representation of FGD two types of brackets for tree linearization:

• ⟨ ⟩ for dependencies
• [ ] for coordination
References

Dependency and non-dependency relations

other non-dependency relations in PDT:
• technical root – effective root of a sentence
• syntactically unclear expressions
  rhematizers; sentence, linking and modal adverbial expressions, conjunction modifiers
• list structures
  names, foreign expressions
• phrasemes