

Corpus Annotation - Sentence and Discourse

1. Underlying layer of syntactic relations and their description in sentence annotation

Corpus linguistics and annotation

Computational Linguistics : what exactly 'computational' means?

- theoretical and applied aspects
- the expansion of the use of computers for linguistic studies based on very large empirical language material → the prevalent use of statistical methods

The appearance of an allegedly new domain, **corpus linguistics**: what is the position of corpus linguistics with regard to computational linguistics

- *computational linguistics*
- *corpus linguistics*
 - the intersection of the two domains is very large
- *theoretical linguistics*

Corpus linguistics and theoretical linguistics

no descriptive framework universally accepted
many different trends in linguistics, **diversity**

→ a highly effective collaboration of researchers needed

it is not appropriate to distinguish between "computational", "corpus" and "real" linguists
discussion on **theoretical** characterization of linguistic phenomena and a **computerized checking** of the adequacy of descriptive frameworks belong to **fundamental goals in linguistics**

Corpora annotation

the requirements of a systematic, intrinsic collaboration (if not a symbiosis) of corpus oriented and computational linguistics with linguistic theory

(a) Many linguists against **statistical methods**, since these may appear as attempts to do without linguistic analyses, using just the outer "brute force"

- b) Other researchers see an attractive goal, or even the center of all appropriate uses of computers in linguistics, in **gathering** large corpora with **searching** procedures
- (c) Still others are aware of the fact that, along with the mentioned goals, there is also the need to use **corpora for theoretical studies**

Annotation of underlying sentence structure

not only to assign part-of-speech (POS) annotations, but also to integrate **syntactic annotations** into the work with large corpora

H. Uszkoreit (2004): time has come for **deep** parsing, and thus, let us add, also for deep corpus annotation

A qualified choice between the existing theoretical approaches (or their parts and ingredients) is necessary to make it possible to use corpora effectively for the aims of theoretical linguistics, as well as of frameworks oriented towards pedagogical and other applications

Dependency vs. constituency

- *dependency grammar*: syntactic relations in the sentence – relation between the *governor* and its *dependent(s)*
- Predicate and its arguments
- K.F.Becker (1837), grammars of German
- Vladimír Šmilauer (1946 and later): the relation between subject and predicate as a special relation
- Lucien Tesnière (1959): actants and circonstants (compare: arguments and adjuncts)

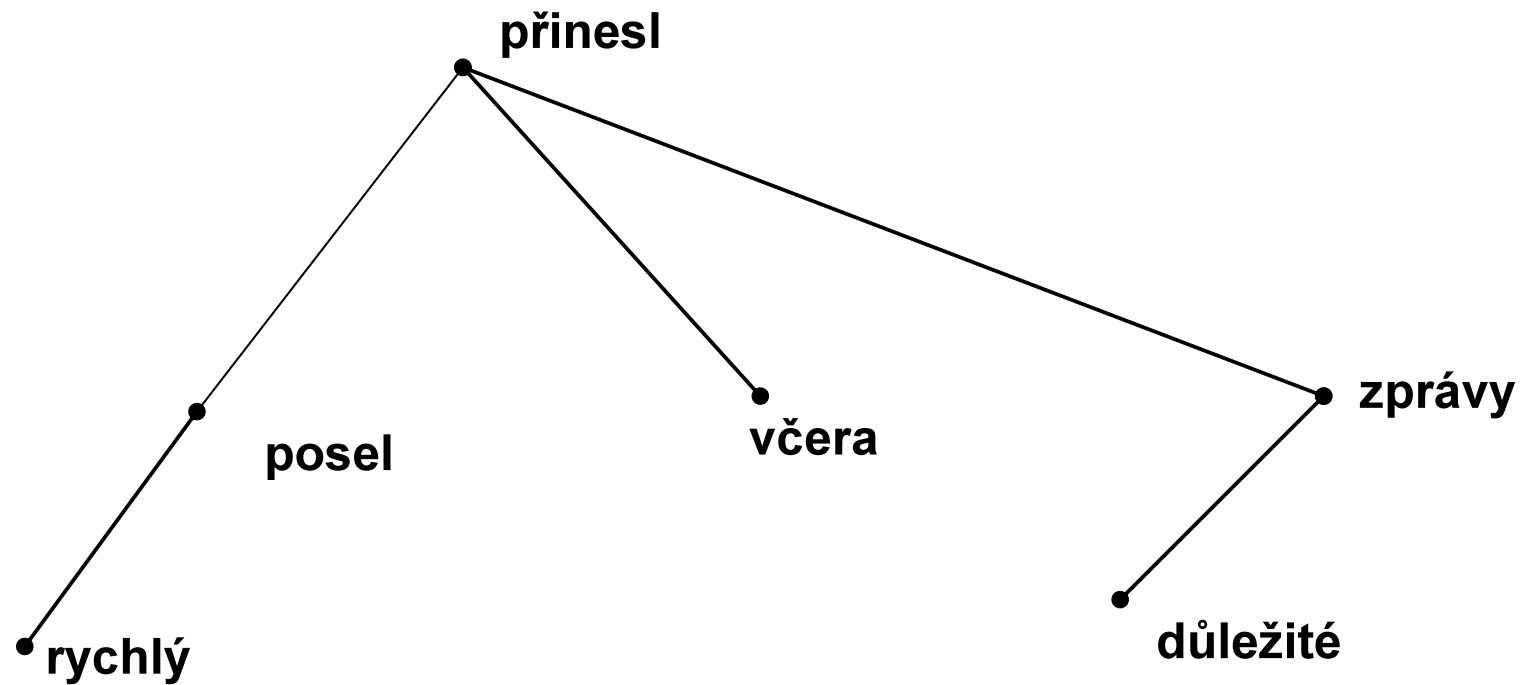
Formal representation

- Syntactic representation of the sentence:
- dependency tree = a graph with a single root, in which every node has a single mother (i.e. the immediately superordinated node) so that there is a single path from each node to the root
binary relation: governor - dependent
- there may be more nodes depending on a single node
- nodes in the tree are ordered both structurally (mother – daughter) and linearly (from the left to the right)

Example of a dependency tree

Rychlý posel přinesl včera důležité zprávy.

Fast messenger brought yesterday important news



Types of dependency, the “direction” of dependency

**Types of dependency (= a set of
dependency relations)**

*Rychlý*_{attr} *posel*_{subj} *přinesl* *včera*_{temp}
*důležité*_{attr} *zprávy*_{obj}

**Which is the governor and which is the
dependent in the given pair?**

verb = the root of the tree

syntactic deletability in endocentric constructions:

(important) news, brought (yesterday)

by analogy: *(messenger) brought (news)*

analogical to: *(messenger) telephoned*

Problems

- Coordination, apposition – a relation of a different kind, “third dimension” (but: Mel’čuk!)
- Condition of projectivity
- ! Phrase structure grammar has the same problems!

Phrase structure, immediate constituents

Criterion: closeness of the relation, binary?

Rychlý posel přinesl včera důležité zprávy.

Fast messenger brought yesterday important news.

(fast messenger)(brought yesterday important news)

((fast)(messenger))

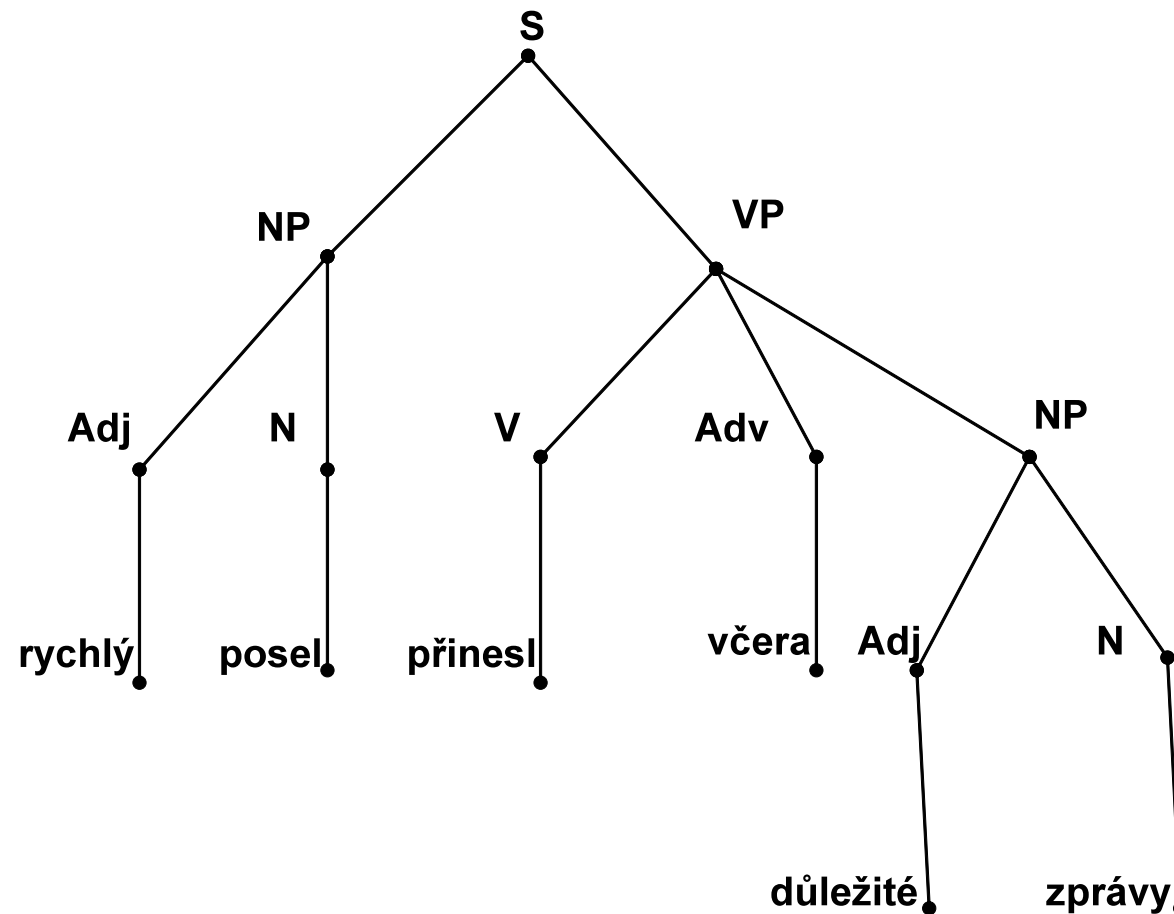
?((brought)(yesterday *important news*))

? ((brought yesterday) (*important news*))

? ((brought) (yesterday) (*important news*))

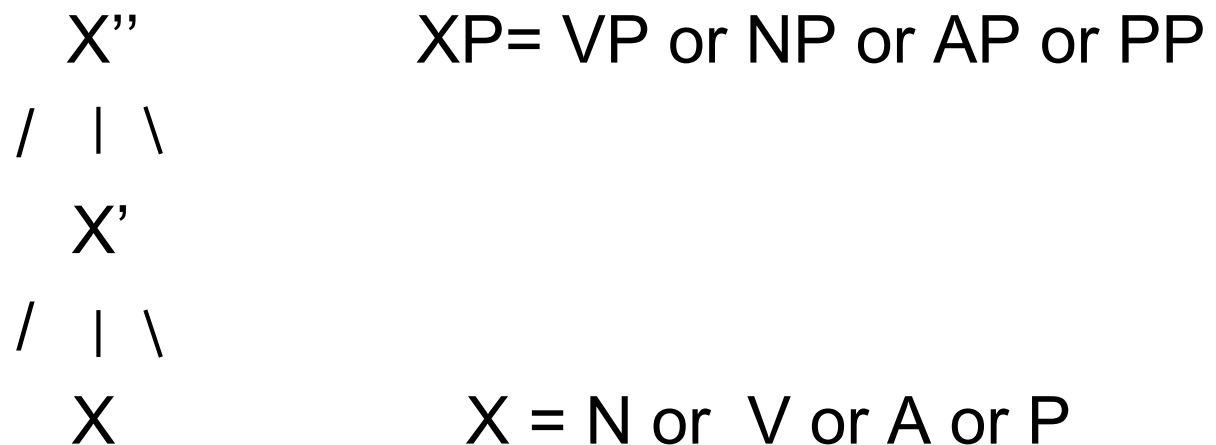
? ((brought) (yesterday) (*important*) (*news*))

Constituent (phrase) structure



Dependency in theoretical description of language

- 'head': Chomsky – *X-bar theory* –
- 4 categories N, V, Adj, Prep as 'heads' of their 'projections' (NP, VP, AP, PP)



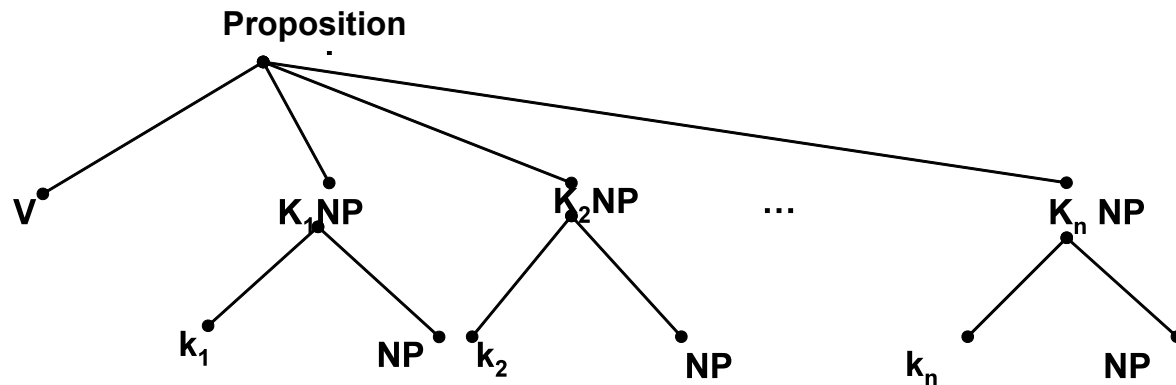
- *Head-driven phrase structure grammars* (Pollard, Sag 1987, 1994)
 - X-bar theory
 - valency
- Lexical-functional grammar LFG (J. Bresnan, 1978; 1982); x Chomsky
 - surface structure: theory X-bar
 - ‘functional’ structure (underlying): predicate and its arguments (verb = the core of the structure)

Case theory as a step towards dependency

- Case grammar: Ch. Fillmore (1964; 1969, etc.)
- ‘case’ = the meaning of morphological case, corresponds to the notion of valency)
- ‘cases’: Agentive, Objective, Addressee, Temporal, Local, ...
- Main motivation: Chomskian ‘deep structure’ does not render semantic relations in a sufficient way

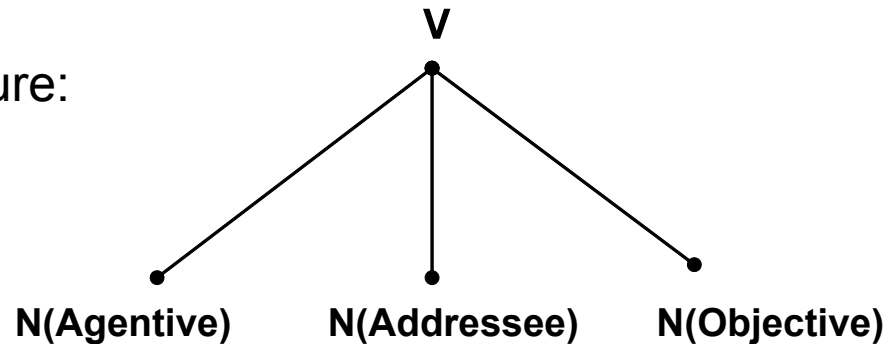
Case theory (2)

- (modality)(proposition)



where k_1, k_2, \dots, k_n = case marker (Agentive, Addressee, Objective, Temporal, Local,...)

- Transition to dependency structure:



Problems of phrase structure

- Ambiguity of information structure – cannot be represented by (standard) constituents (Hajičová-Sgall, 1975)

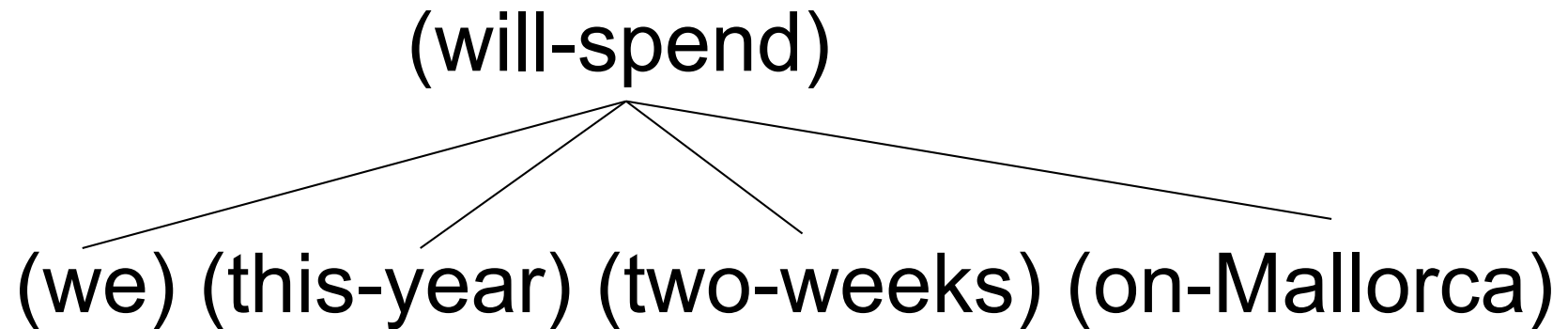
This year we will spend two weeks on Mallorca.

(How will you spend your holidays this year?)

two weeks on Mallorca – is not a phrase

- in other contexts:
(How will you travel this year?) ... *we will spend two weeks on Mallorca.*
(Where will you spend two weeks this year?)... *on Mallorca.*

No problems with dependency



Possible solution for phrase structure

- M. Steedman (1996, 2000, 2002): introduces ‘non-standard constituents’:
- a ‘floating’ border line between constituents - makes it possible to capture:
 - structural ambiguity
 - different prosody according to the articulation of the sentence into its topic and focus
 - *Fred | ate | the BEANS.*
 - *FRED | ate the beans.*
 - *Fred ATE | the beans.*

Praguian view

- (a) Syntactic dependency is handled as a set of relations between **head words and their modifications** (arguments and adjuncts)
 - the relations of coordination (conjunction, disjunction and other) and of apposition, understood as relations of a “further dimension”
 - the tectogrammatical representations are more complex than mere dependency trees
- (b) the topic-focus articulation (**information structure**) of sentence
 - communicative dynamism (underlying word order)
 - the dichotomy of contextually bound (CB) and non-bound (NB) items

Tectogrammatical representations

the core of a tectogrammatical representation: a **dependency** tree with the **verb as its root**

direct dependents are **arguments**: Actor, Objective (Patient), Addressee, Origin and Effect, and adjuncts (of location and direction, time, cause, manner, and so on)

no nodes corresponding to function words (prepositions, auxiliary verbs, articles) or to grammatical morphs – their correlates: indices of node labels

values of morphological categories (tense, number, and so on) have the form of indices, grammatemes

The condition of projectivity

projective dependency trees:

for every pair of nodes in which a is a rightside (leftside) daughter of b , every node c that is less (more) dynamic than a and more (less) dynamic than b depends directly or indirectly on b

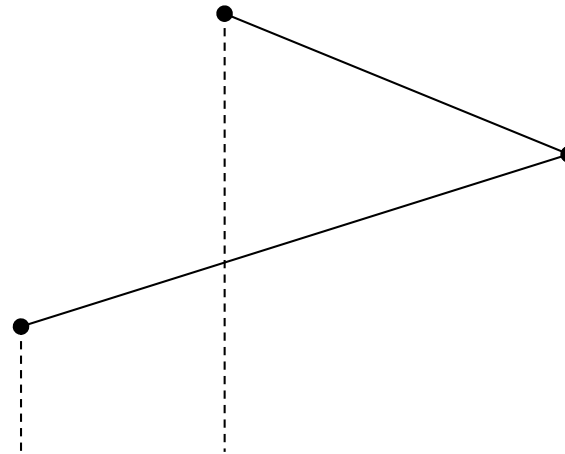
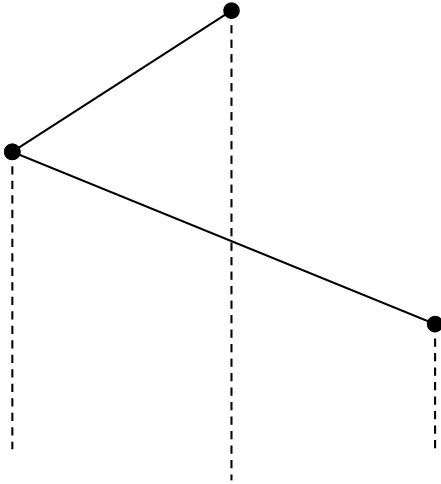
(*indirectly* = transitive closure)

Special treatment: coordination and apposition

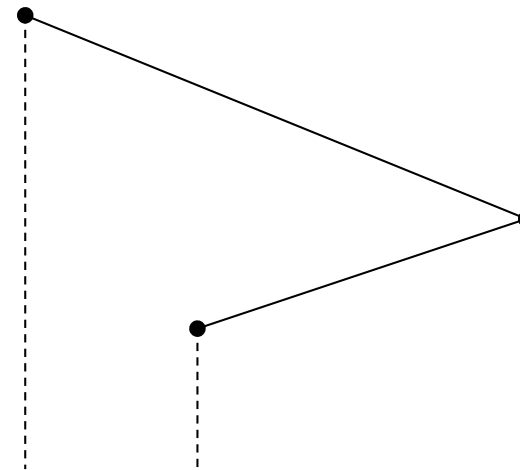
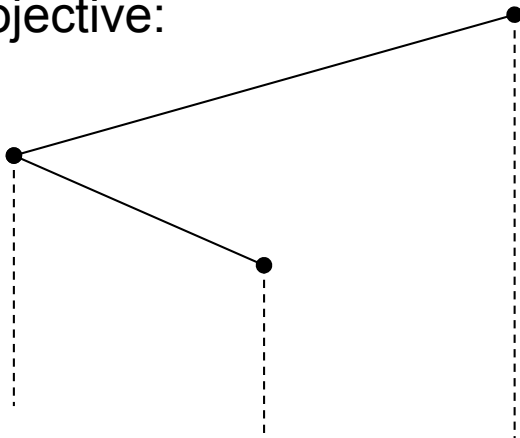
projective trees thus come relatively **close to linear strings**; they belong to the most simple kinds of patterning

Condition of projectivity

non-projective :



projective:



Prague Dependency Treebank

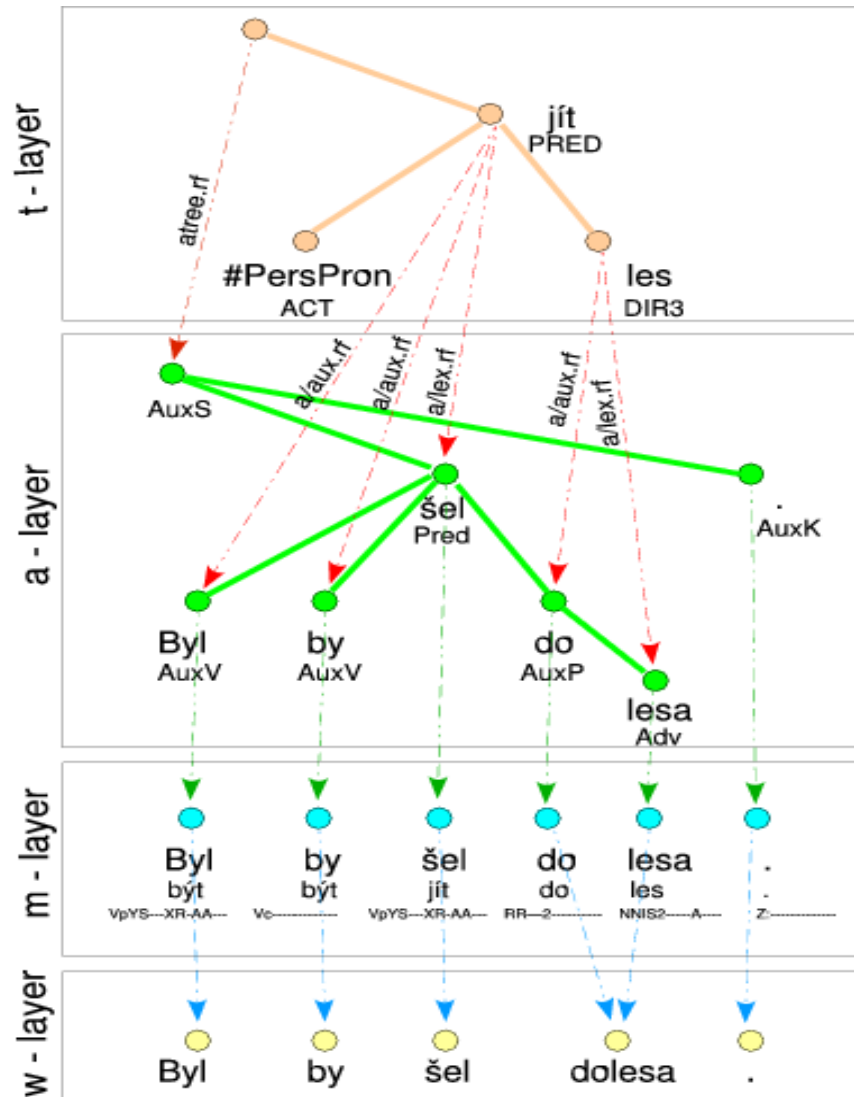
an annotated collection of Czech texts, randomly chosen from the Czech National Corpus (CNK), with a mark-up on **three layers**:

- (a) morphemic
- (b) surface shape “analytical”
- (c) underlying (tectogrammatical)

the current version annotated on all three layers (<http://ufal.mff.cuni.cz/pdt2.0>, with the data themselves available at LDC under the catalog No. LDC2006T01):

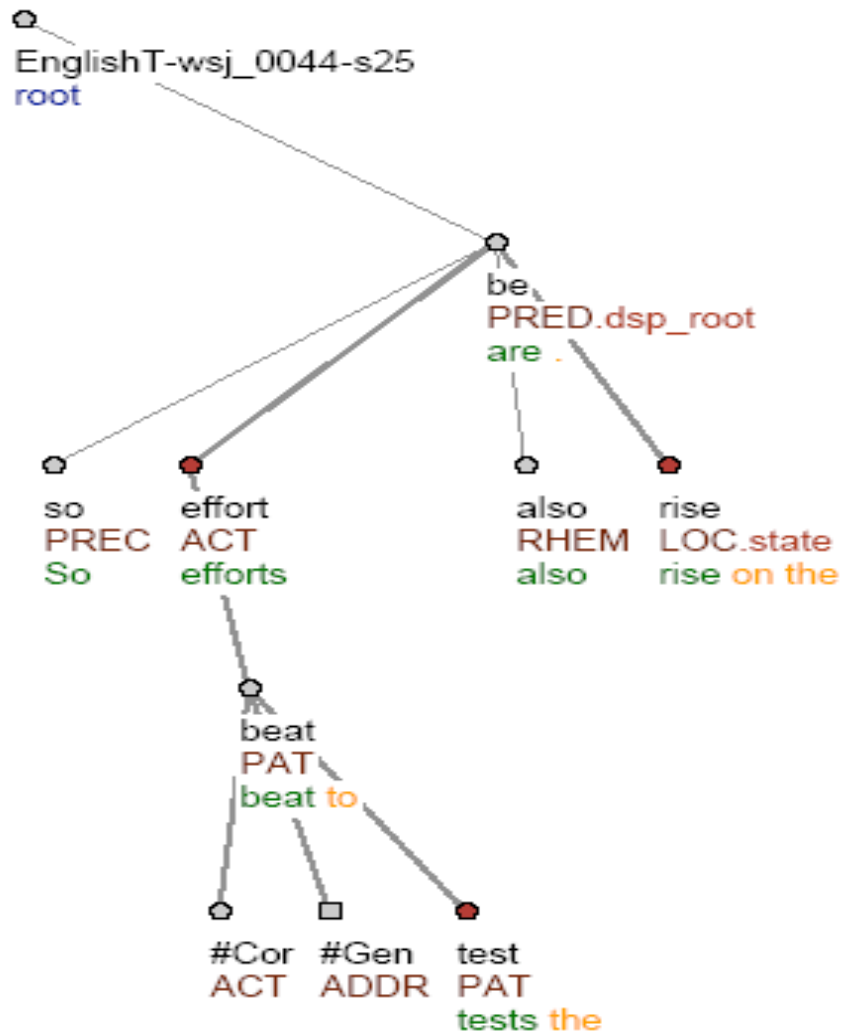
3165 documents (mainly from journalistic style)
comprising 49431 sentences and 833195
occurrences of tokens (word forms and punctuation marks)

Prague Dependency Treebank:



- 4 layers (3 annotation layers + raw) in the Prague Dependency Treebank
- T-layer (deep syntax or tectogrammar) contains some discourse relations already
- 1. coordinations
- 2. subordinate (dependent) clauses
- 3. NO inter-sentential marking (yet)
- some of the discourse connectives marked by the label PREC

An example of a dependency tree



File: wsj_0044.t.gz, tree 25 of 135

“ So efforts * to beat the tests are also on the rise. ”
Takže se častěji objevují snahy zvítězit nad testy.

Tectogrammatical tree structures in the PDT

every **node** of the tectogrammatical representation (TGTS, a dependency tree) has a label consisting of:

- the *lexical value* of the word,
- its (*morphological*) *grammatemes* (i.e. values of morphological categories),
- *functors* (with a more subtle differentiation of syntactic relations by means of *subfunctors*, e.g. 'in', 'at', 'on', 'under')
- the topic-focus articulation (TFA) attribute containing values for *contextual boundness*
- some basic *coreferential* links (including intersentential ones)
- in case of *surface deletions* TGTSs may contain nodes not present in the morphemic form of the sentence

in annotating texts from the Czech National Corpus: **some specific deviations** from theoretically conceived TRs:

the most important deviation: the tectogrammatical tree structures of PDT have the form of trees **even in cases of coordination**
→ the coordinating conjunctions are handled as specific nodes (with a specific index)

Dependency relations in PDT

- Valency theory
- Distinction between inner participants (arguments) and free modifications (adjuncts)
 - Arguments: Agent, Patient, Addressee, Origin, Effect
 - Adjuncts: temporal, locative, directional, causal, instrument, means, ...

Arguments vs. adjuncts

- (a) Do the rules of the language allow for the occurrence of the modification in question with every verb?
- (b) Is the modification allowed to occur more than once while depending on a single verb token?
- Arguments: (a) no (except for Agent)
(b) no
- Adjuncts: (a) yes (exceptions listed)
(b) yes

- Distinction between semantically obligatory and semantically optional
- Dialogue test:
- *arrive* – *Speaker: He has arrived.*
A: When? S: I do not know.
A: Where to? S: I do not know.*

Requirements on lexical entries

- Frames:
- Arguments: listed, marked as obligatory or optional
- Adjuncts: need not be listed, only exceptions and obligatoriness (*to behave HOW, to stay WHERE, ...etc.*)