Dependency Grammars and Treebanks:

Introduction

Daniel Zeman, Jiří Mírovský

(slides credit: Markéta Lopatková)

ÚFAL MFF UK

{zeman,mirovsky}@ufal.mff.cuni.cz

Dependency Grammars and Treebanks (NPFL075)

Thursday, SU1, 12:20-13:50

Alternating weeks with lectures vs. practical sessions

Lectures: Daniel Zeman

Practicals: Jiří Mírovský

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

Requirements:

- Homework
- Activity
- ~~Final test~~

Assessment:

excellent (= 1)	≥ 90%
very good (= 2)	≥ 70%
good (= 3)	≥ 50%

Dependency Grammars and Treebanks

Treebank as a collection of:

- linguistically annotated data
- tools and data format(s)
- documentation

⇨

- Family of Prague Dependency Treebanks (PDT, PCEDT)
- Universal Dependencies
- HamleDT, PropBank, …

Another point of view:

- underlying linguistic theory
- annotation scheme
- framework for annotation of different languages

Outline of the lecture

- Introduction: dependency grammar in a nutshell
- Tree-based structures informally
    - phrase structure / constituency trees
    - dependency trees
- How to detect a dependency relation?
- A bit of math …
- Problem with free word order

Dependency grammar (DG)

notion of DG in a nutshell:

The dependency grammar is

- a model developed by Lucien Tesnière (1893-1954) and
- based on structuralism
- to describe the syntax of natural languages.

The main concern of the dependency grammar is

- the description of the dependency structure of a sentence, i.e. the structure of dependency relations between the elements of a sentence.

Dependency grammar (DG)

notion of DG in a nutshell:

The dependency grammar is

- a model developed by Lucien Tesnière (1893-1954) and
- based on structuralism
- to describe the syntax of natural languages.

The main concern of the dependency grammar is

- the description of the dependency structure of a sentence, i.e. the structure of dependency relations between the elements of a sentence.

- dependency as an asymmetric binary relation between language units
- governing/modified unit (head) – dependent/modifying unit (modifier)
  → word (morph) grammar … *lexicalization*
  → no phrase nodes
- dependency trees, with edges ~ *dependency relations* (mostly)

A bit of history ….

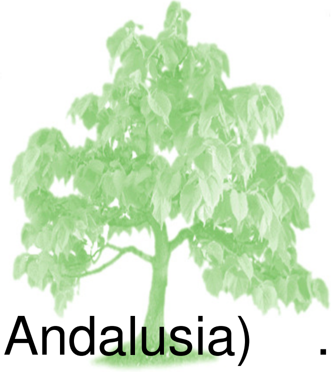
structural linguistics:

(based on Ferdinand de Saussure: *Course in General Linguistics*, 1916)

- *synchronic* approach (vs. diachronic)
- *sign*: “signified” (idea, concept) – „signifier“ (means of expressing)
- examining language as a (static) system of interconnected units
- stress on structure (signs cannot be examined in isolation)
- *syntagmatic* vs. *paradigmatic* relations
- *langue* (idealized abstraction of language) vs. *parole* (language as actually used)

structuralist schools:

- Genova School (course 1909-1912): Ferdinand de Saussure, Albert Sechehaye, Charles Bally
- Prague School (1926–1939): Vilém Mathesius, Bohumil Trnka, Bohuslav Havránek, Jan Mukařovský, Roman Jakobson, Nikolai Trubeckoj, Sergej Karcevskij
- Copenhagen School (1930-1950): Louis Hjelmslev, glossematics
- “American structuralism” (1920-50): Leonard Bloomfield, Charles Hockett



Dependency-based approaches

- Pāṇini (6th–4th century BC; India); Ibn Maḍāʾ(12th century AD; Andalusia) . the term *dependency*

- Franz Kern and others († 1894, esp. pedagogy) ... sentence diagrams
- Lucien Tesnière (1893–1954; France) … valency, “stemma” (unordered)

motivation for current computational linguistics / NLP:

- David Hays (1950–1960, machine translation ru→en)
- Zellig H□□□is (si□ce 1930, † 1992; li□guistics □s □pplie□ m□them□tics; methodology of linguistic analysis)
- *Dependenzgrammatik* … esp. Jürgen Kunze (from 1960s, 1975)
  Valenzgrammatik … esp. Gerhard Helbig (from 1960s)
- Richard Hudson (from 1970s, 1984) … *Word Grammar*
- Michael Halliday … *Systemic Functional Grammar*

Dependency-based approaches (cont.)

- *Meaning-Text Theory* (MTT) … applied esp. in machine translation, lexicography; Igor Meľčuk, Aleksandr Žolkovskij (1965-)
- *Functional Generative Description* (FGD) ... applied in treebanks from the Prague dependency family, used esp. for machine translation; Petr Sgall and his school (1967-)
- *Universal Dependencies* (UD) … since 2014, Joakim Nivre et al.

Corpora with dependency trees

- PropBank (1995)
  http://propbank.github.io/
- Prague dependency treebank (1996) first Czech, then Arabic, English, …
  http://ufal.mff.cuni.cz/pdt.html
- HamleDT project (from 2012) http://ufal.mff.cuni.cz/hamledt
- Universal Dependencies (from 2013) http://universaldependencies.org/

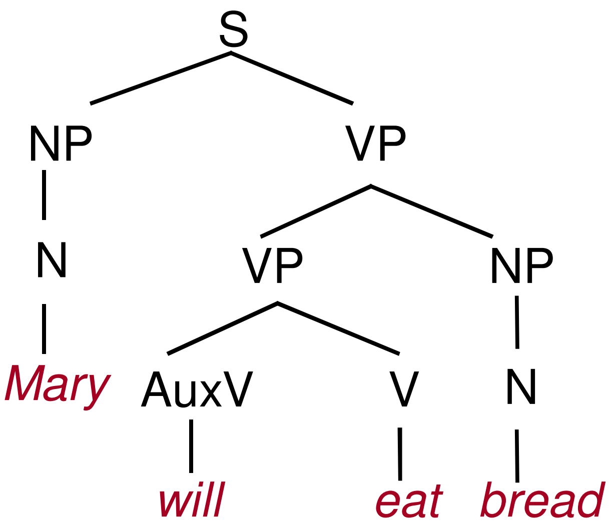
- Danish Dep. Treebank
  http://mbkromann.github.io/copenhagen-dependency-treebank/
- Finnish: Turku Dependency Treebank
  http://bionlp.utu.fi/fintreebank.html
- Negra corpus
  http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html
- TIGERCorpus
  http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html/
- SynTagRus Dependency Treebank for Russian

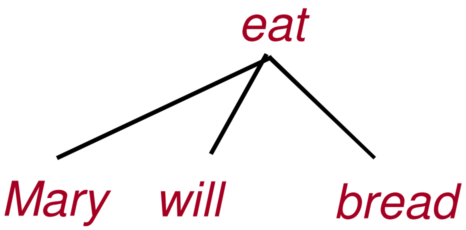
Outline of the lecture

- Introduction: dependency grammar in a nutshell
- Tree-based structures informally
    - phrase structure / constituency trees
    - dependency trees
- How to detect a dependency relation?
- A bit of math …
- Problem with free word order

Phrase structure vs. dependency tree

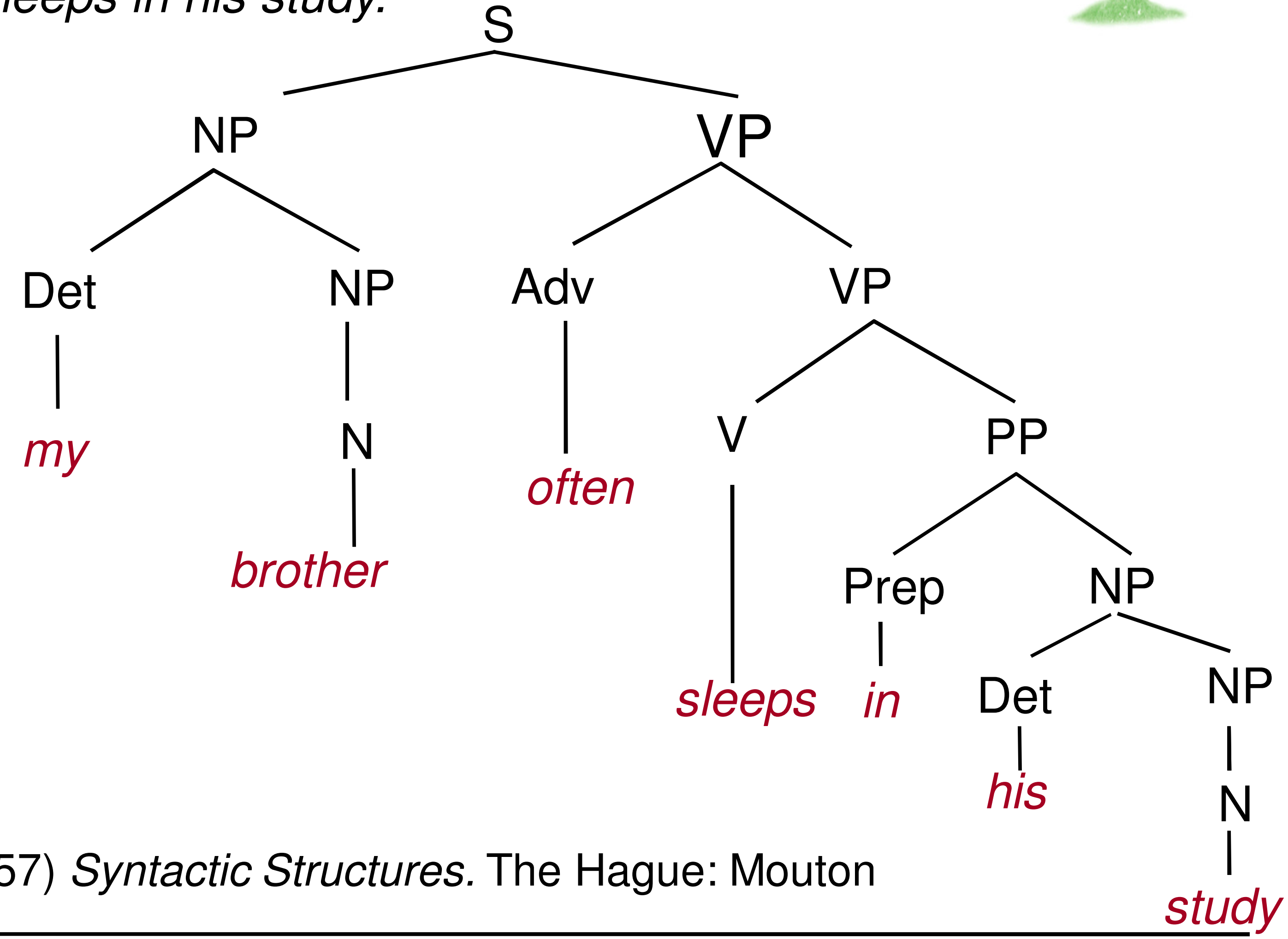
Mary will eat bread.





Phrase structure trees

My brother often sleeps in his study.

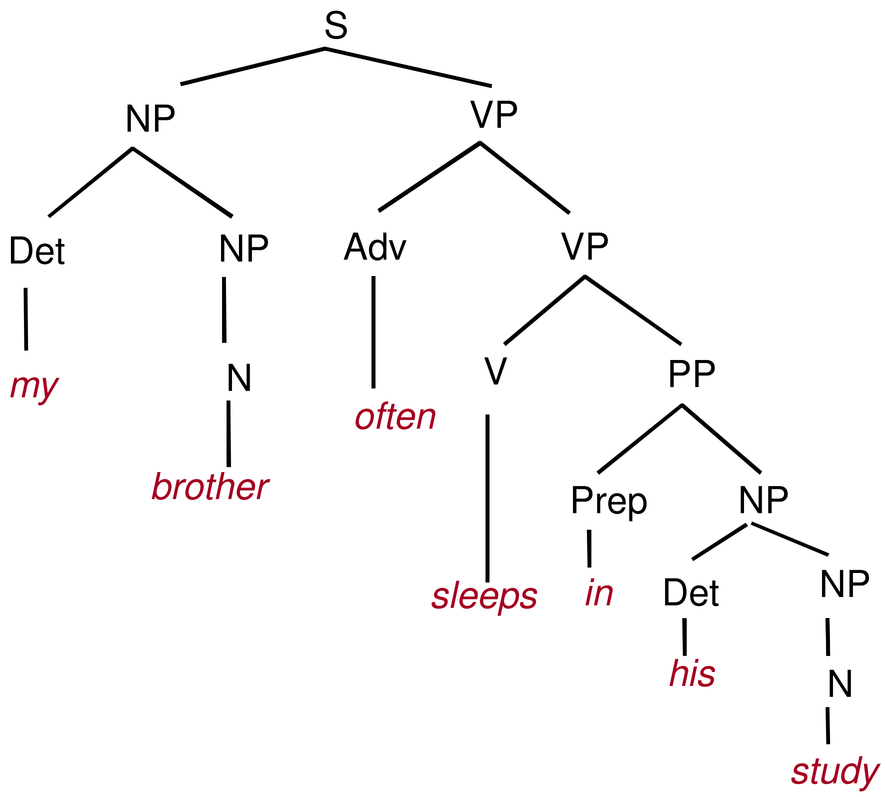


Noam Chomsky (1957) *Syntactic Structures.* The Hague: Mouton

Phrase structure trees (cont.)

Pros

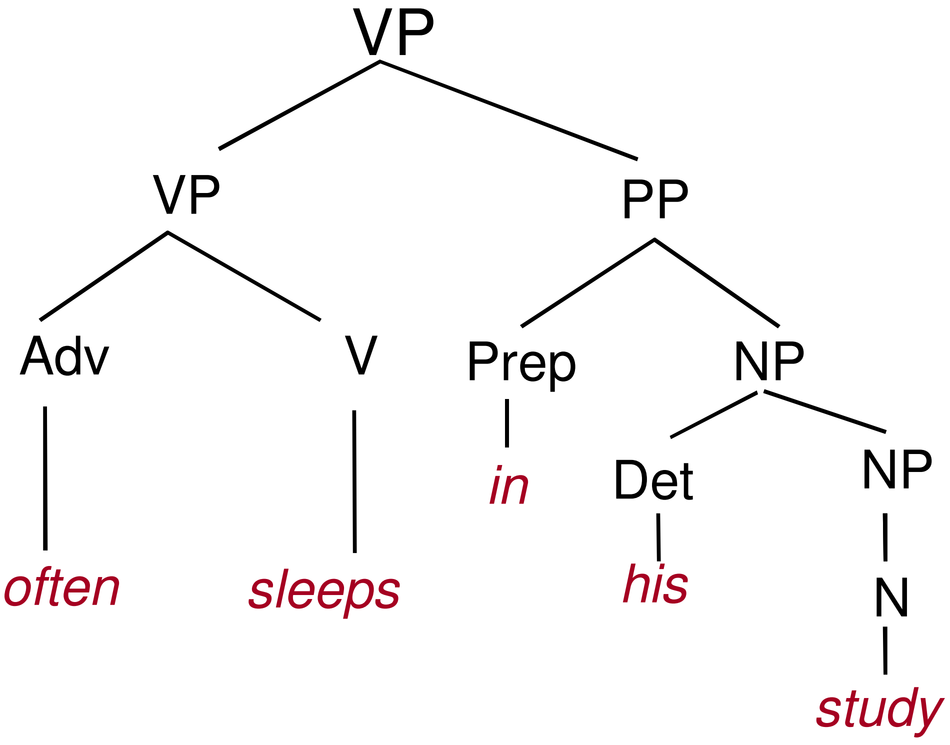
- rich syntactic structure
- derivation history / ‘closeness’ of a complementation
- CFG-like
- coordination, apposition
- derivation of a grammar



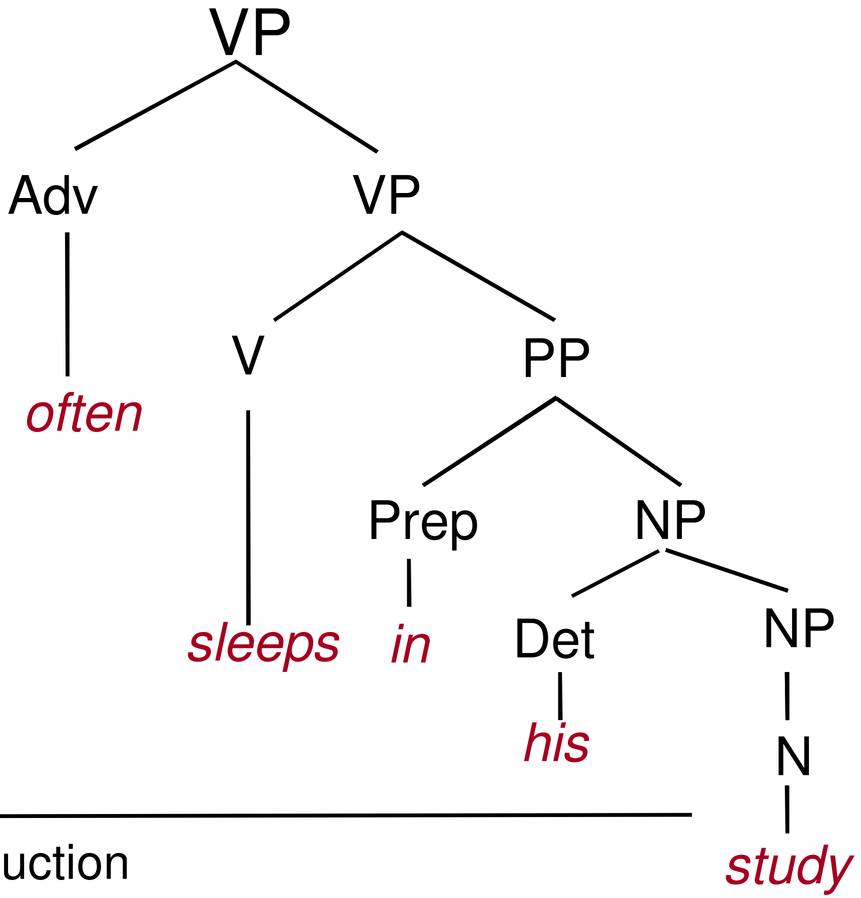
Phrase structure trees (cont.)

derivation history / ‘closeness’:

… often sleeps in his study



… often sleeps in his study

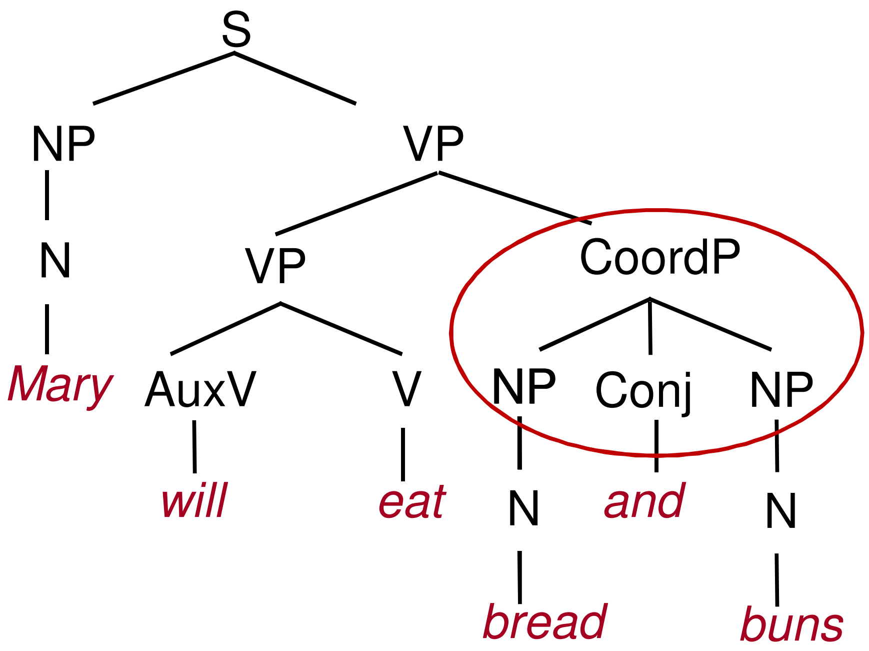


Phrase structure trees (cont.)

Pros

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

(BUT is it appropriate?)



Phrase structure trees (cont.)

Pros

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

  (BUT is it appropriate?)

Contras

- complexity

  (number of non-terminal symbols)
- secondary predicates

  (‘two dependencies’)

  přiběhl bos [(he) arrived barefooted]

  She declared the cake beautiful.
- free word order

  discontinuous ‘phrases’

  non-projectivity
- binary division of a clause

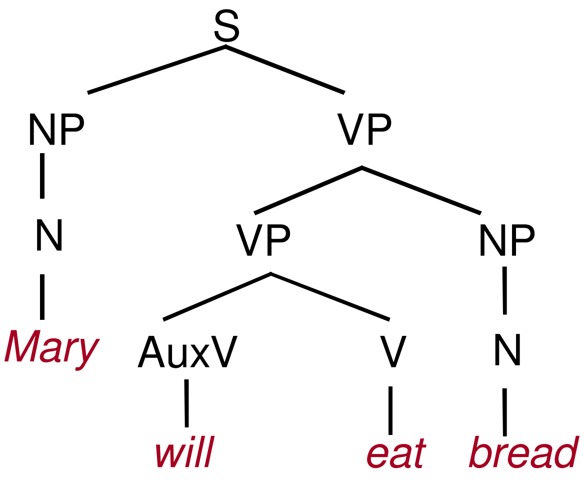
  (imposed by logic,

  not by language structure)

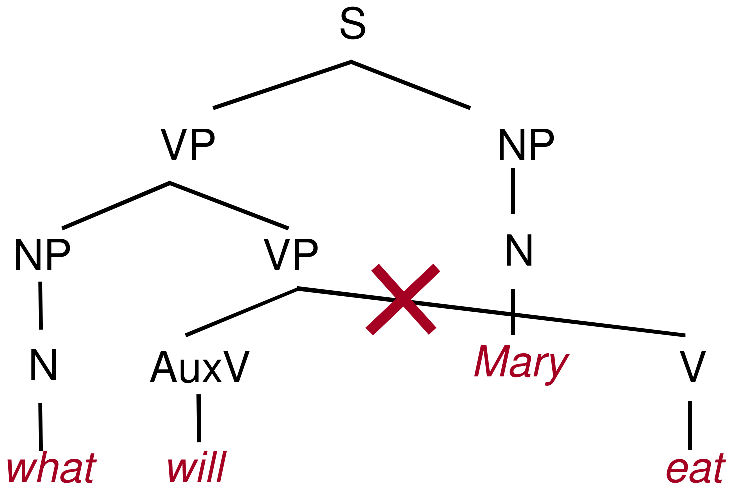
Phrase structure trees (cont.)

discontinuous ‘phrases’: solution for English

Mary will eat bread.



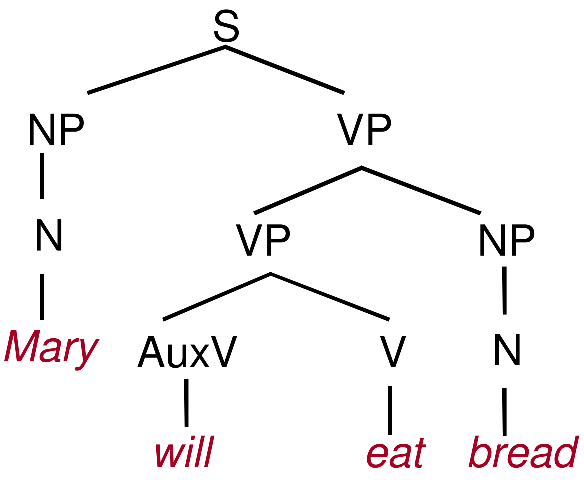
What will Mary eat?



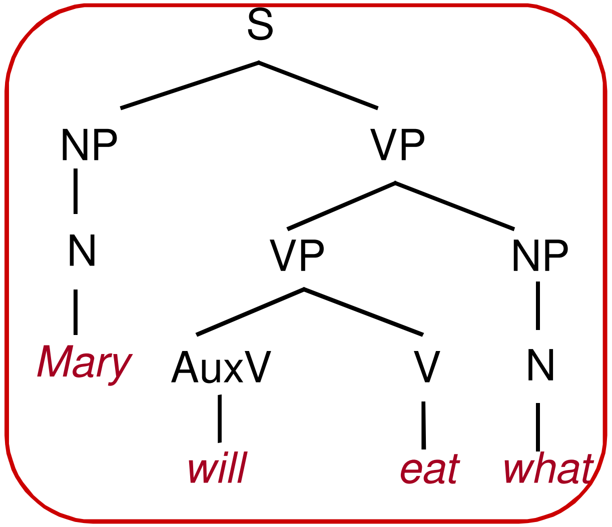
Phrase structure trees (cont.)

discontinuous ‘phrases’: solution for English

Mary will eat bread.



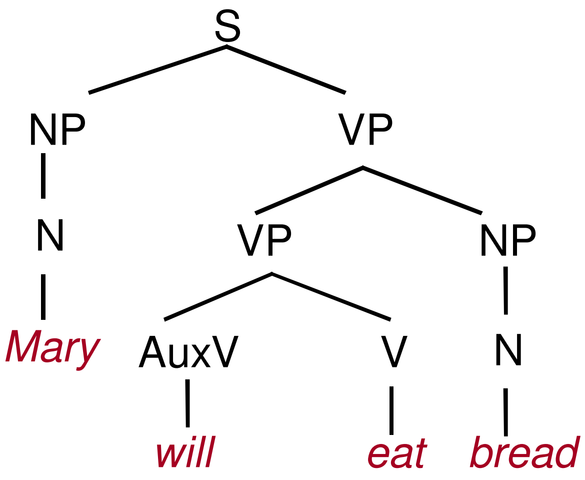
What will Mary eat?



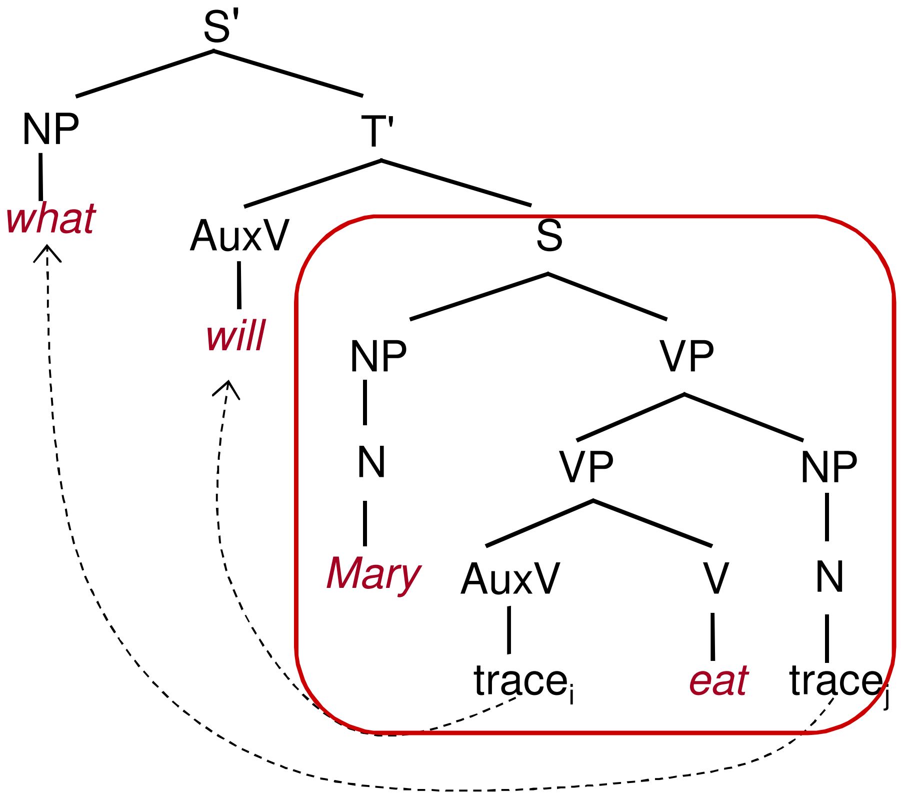
Phrase structure trees (cont.)

discontinuous ‘phrases’: solution for English

Mary will eat bread.



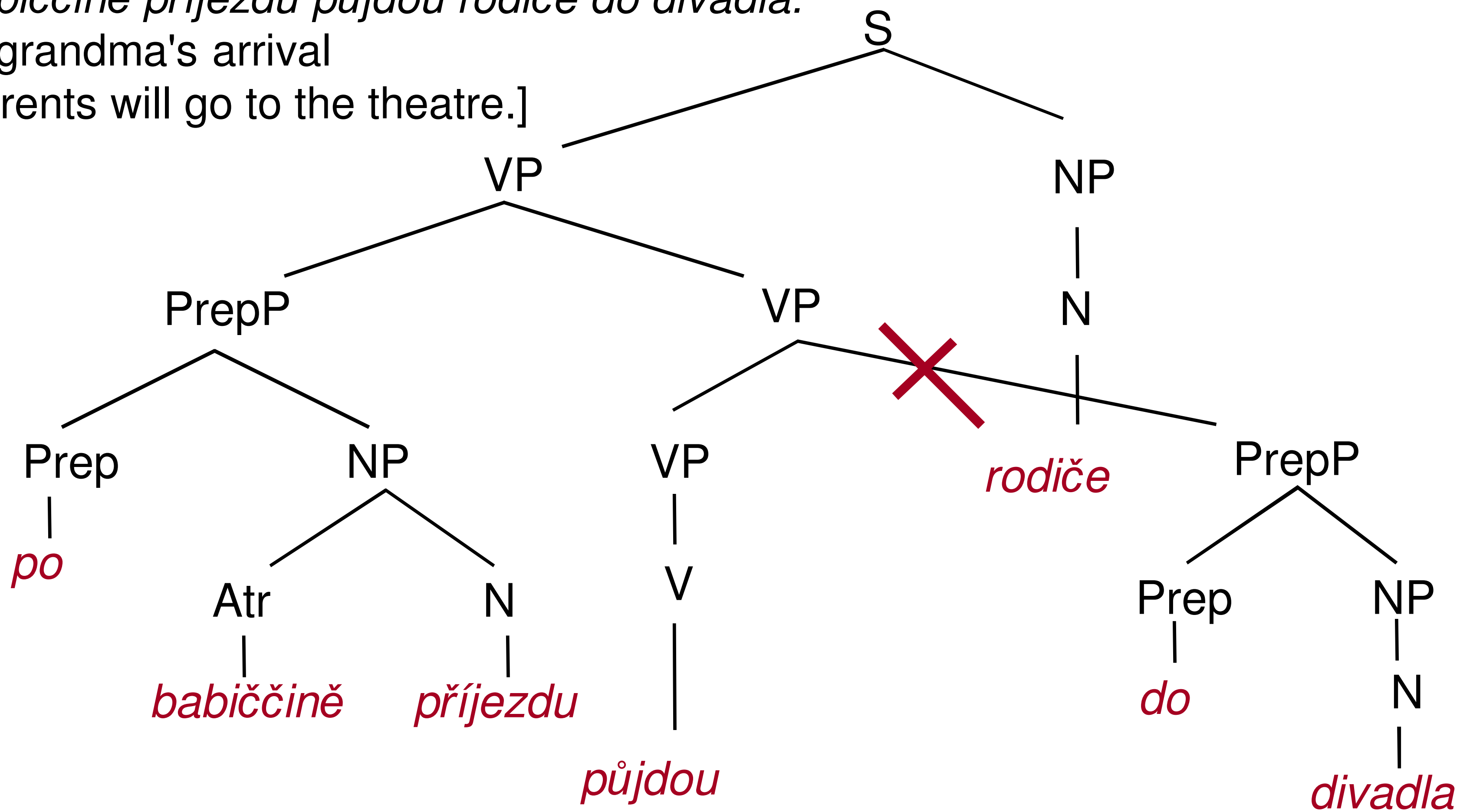
What will Mary eat?



Phrase structure trees (cont.)

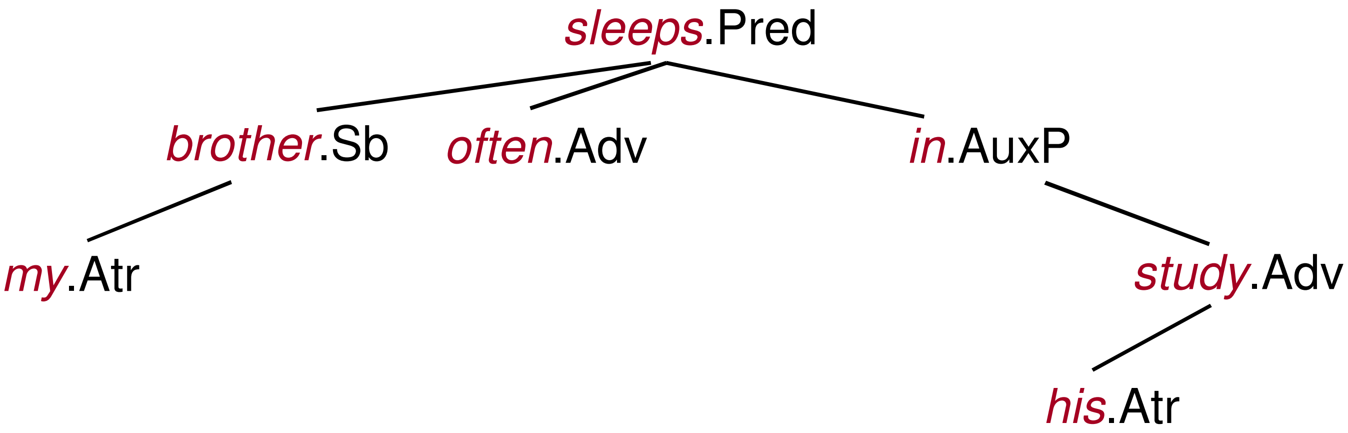
discontinuous ‘phrases’:

Po babiččině příjezdu půjdou rodiče do divadla.
[After grandma's arrival
the parents will go to the theatre.]



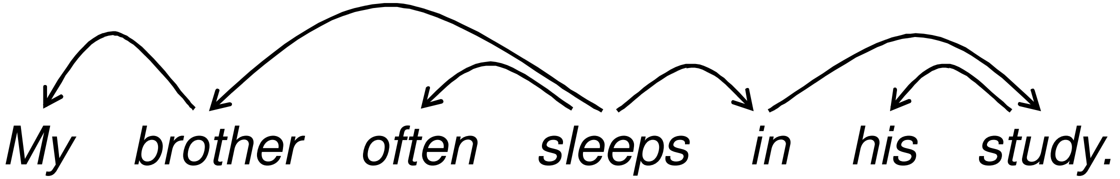
Dependency trees

My brother often sleeps in his study.



Lucien Tesnière (1959) *Éléments de syntaxe structurale.* Editions Klincksieck.

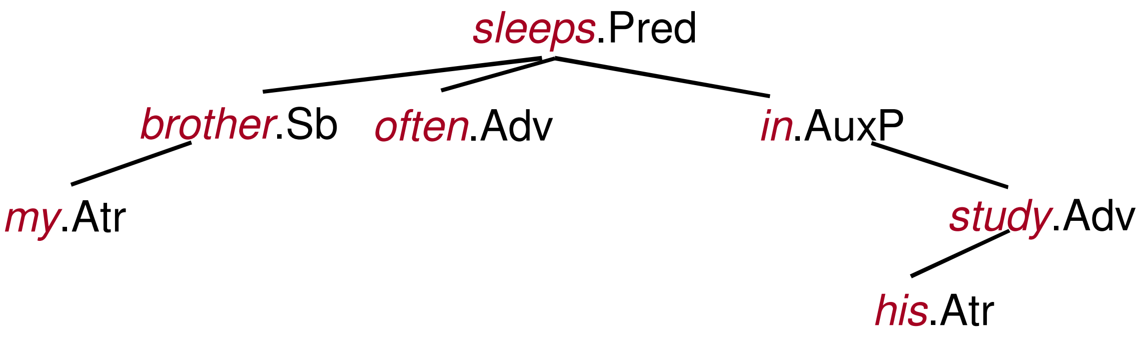
Igor Mel'čuk (1988) *Dependency Syntax: Theory and Practice.* State University of New York Press.



Dependency trees

Pros

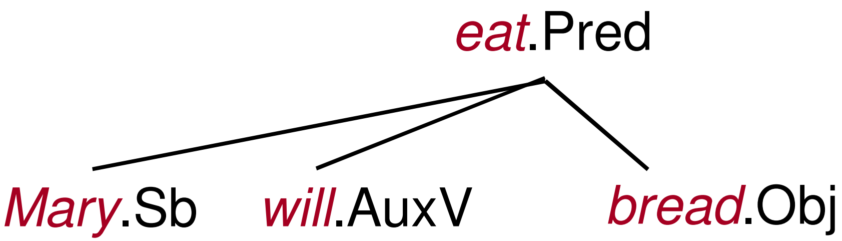
- economical, clear
  (complex labels, ‘word’~ node)
- head of a phrase
- free word order



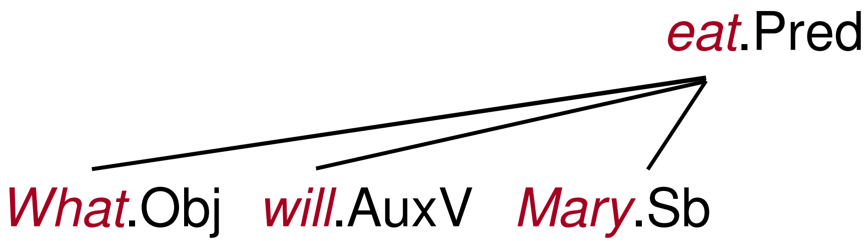
Dependency tree

discontinuous ‘phrases’: no problem

Mary will eat bread.

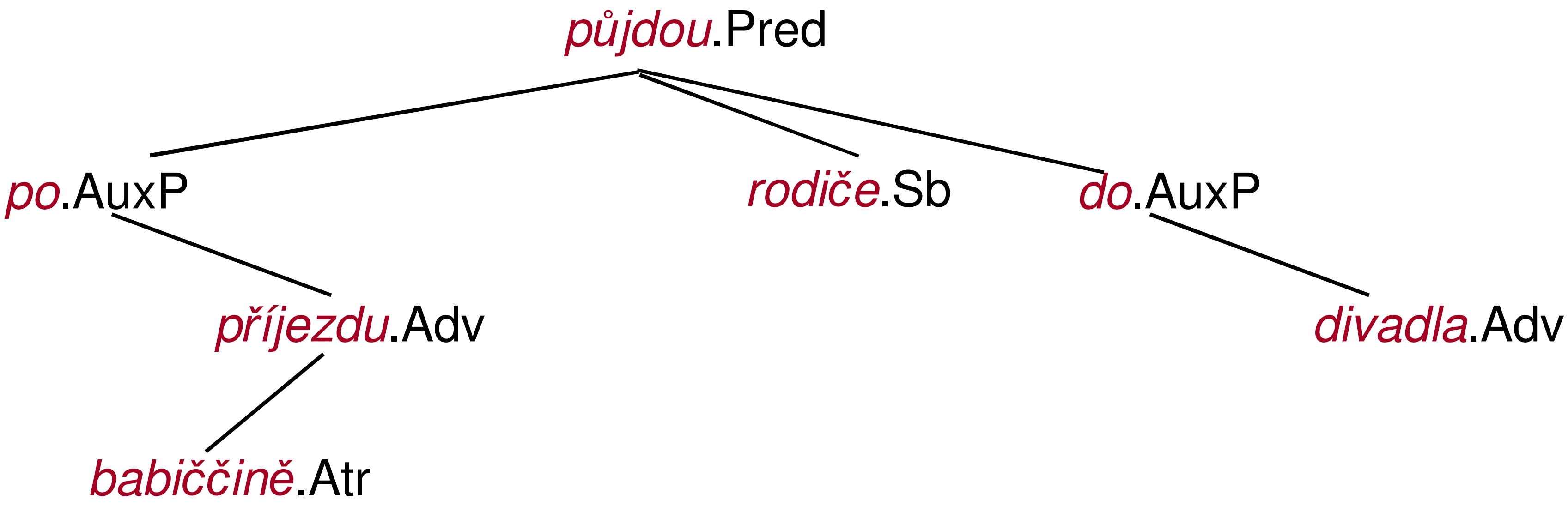


What will Mary eat?



Dependency tree

Po babiččině příjezdu půjdou rodiče do divadla.
[After grandma's arrival the parents will go to the theatre.]



Dependency trees

Pros

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

Contras

- no derivation history / 'closeness'
- coordination, apposition
- secondary predicates (‘two dependencies’)

  přiběhl bos [(he) arrived barefooted]
  She declared the cake beautiful.

Outline of the lecture

- Introduction: dependency grammar in a nutshell
- Tree-based structures informally
    - phrase structure / constituency trees
    - dependency trees
- How to detect a dependency relation?
- A bit of math …
- Problem with free word order

Dependency Relations

- *semantic dependencies* … semantic predicates and their argu

  cf. *Sam likes Sally.* …. like(Sam, Sally)

  vs. *new car* (= car being new) … New(car)

- *syntactic dependencies*

Dependency Relations (cont)

- *morphological dependencies* (agreement)

  cf. *Mary comes here.* vs. *Children come here*.
  cf. *this house* vs. *these houses*
  cf. *strom je zelený* ‘house is green-sg-inan’
  vs. *stromy jsou zelené* ‘houses are green-pl-inan’
  vs. *mužíci jsou zelení* ‘men are green-pl-anim’

- *intra-word dependencies* (→ derivational morphology)
- *prosodic dependencies*

  cf. *clitic* (a syntactically autonomous unit prosodically dependent on its host)
  He’ll stop. There’s a problem. Peter’s hat. …
  -li; jsem, jsi …, bych, bys, …. ; se, si, mi, ti, mu, mě, tě, ho, …(tam, tu; však, tak)

Syntactic Dependency Relations

- dependency as an asymmetric binary relations between language units

  → detecting heads: not commonly agreed criteria
- number of linguistic criteria
  - e.g., verb as a syntactic center of a sentence
- BUT treebanks:

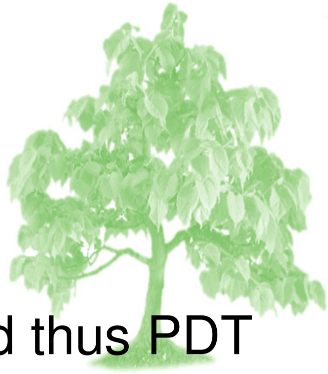
  annotation schemata reflect technical considerations
  - tree-based data format
  - 1:1 correspondence between nodes and tokens

Detecting Syntactic Dependencies I

possible reduction criterion … FGD (Sgall et al. ,1986), and thus PDT

- “dependent member of the pair may be deleted
  while the distributional properties are preserved”
  (→ correctness is preserved)
- endocentric constructions

e.g.	*malý stůl* → *stůl*	*small table* → *table*
	přišel včas → *přišel*	*he came in time* → *he came*
	(přišel) velmi brzo → *(přišel) brzo*	*(he came) very* soon → *(he came) soon*



Detecting Syntactic Dependencies I

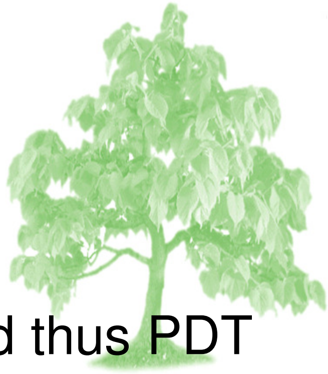
possible reduction criterion … FGD (Sgall et al. ,1986), and thus PDT

- “dependent member of the pair may be deleted
  while the distributional properties are preserved”
  (→ correctness is preserved)
- endocentric constructions
- exocentric constructions … *principle of analogy* (delexicalization)

Prší. [(It) 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

Detecting Syntactic Dependencies I



possible reduction criterion … FGD (Sgall et al. ,1986), and thus PDT

- “dependent member of the pair may be deleted
  while the distributional properties are preserved”
  (→ correctness is preserved)
- endocentric constructions
- exocentric constructions … *principle of analogy* (delexicalization)

Prší. [(It) 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

- plus technical considerations (compare also with “the school grammar”)

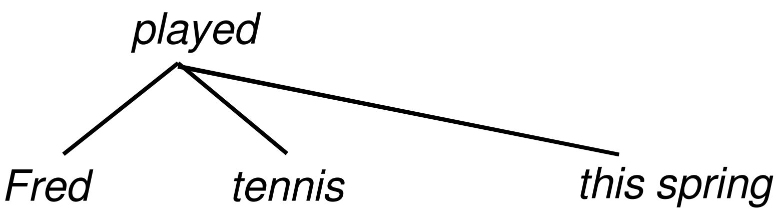
e.g.: prepositions are below nouns;
auxiliary verbs are (typically) below content verbs

Detecting Syntactic Dependencies II

constituent-based criteria (Osborne, 2019)

- each complete subtree must be a “constituent”, based of formal tests, esp:
  - topicalization — *permutation test*
  - clefting and pseudoclefting
  - proform substitution (replacement) — *proform tests*
  - answer fragments
  - coordination

Fred played tennis this spring.



Topicalization:
*… but **tennis** Fred did play this spring.*
***This spring** Fred played tennis.*

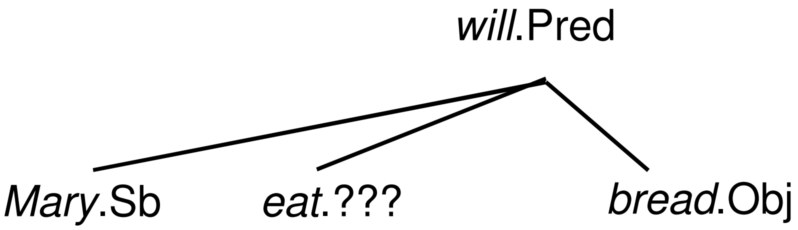
Clefting:
*It was **Fred** who played tennis this spring.*
*It was **tennis** that Fred played this spring.*
*It was **this spring** that Fred played tennis.*

Detecting Syntactic Dependencies II

<span style="color:red">*constituent-based*</span> criteria (Osborne, 2019)

- BUT: applied also for (more-or-less) technical solutions

*Mary **will eat** bread..*



⇨ lexical verb should be a dependent

Topicalization:
*… and **eat** Mary certainly will.*

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

Answer fragment:
What will Mary do? Eat.

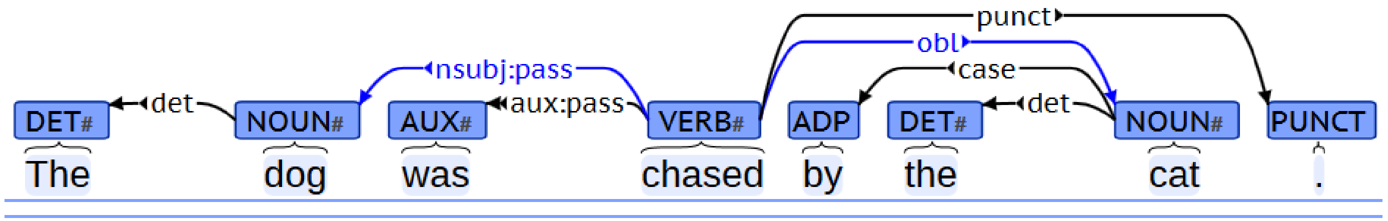
VP-ellipsis:
Peter will eat and Mary will, too.

Detecting Syntactic Dependencies III

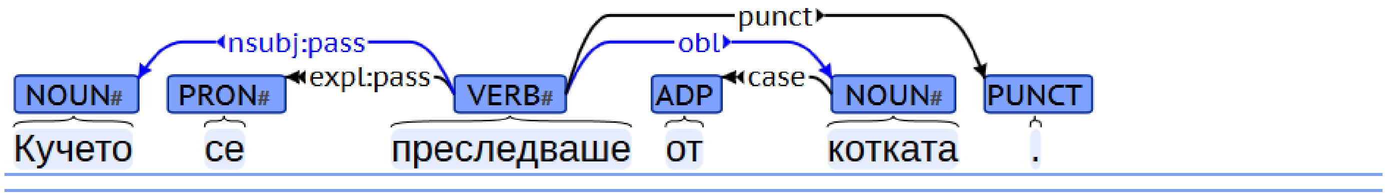
criterion of *maximal parallelism* between languages

… Universal Dependencies

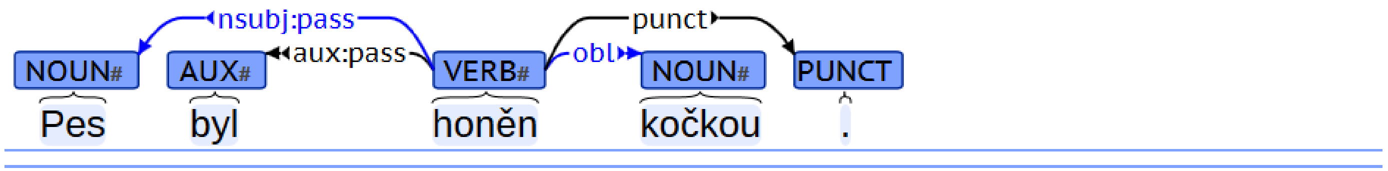
English:


punct
obl
case
nsubj:pass
det
aux:pass
det
DET NOUN AUX VERB ADP DET NOUN PUNCT
The dog was chased by the cat .

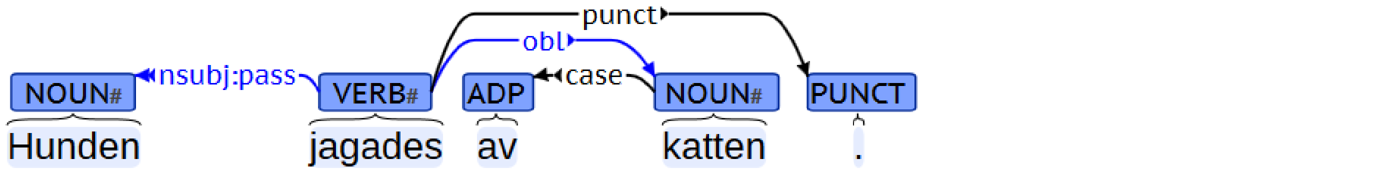
Bulgarian:


punct
obl
nsubj:pass
expl:pass
case
NOUN PRON VERB ADP NOUN PUNCT
Кучето се преследваше от котката .

Czech:



Swedish:

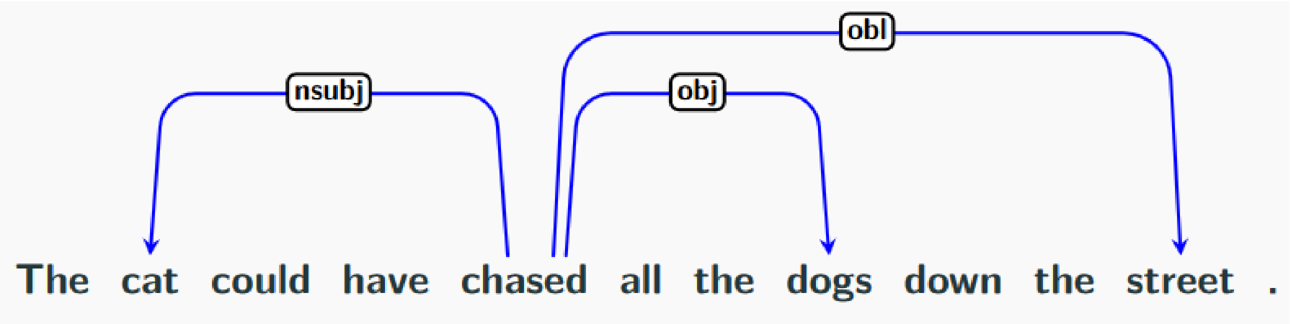


Detecting Syntactic Dependencies III

criterion of *maximal parallelism* between languages

… Universal Dependencies

- the upper levels of UD trees should be as similar as possible across languages
    - dependency relations hold primarily between content words (rather than being indirect relations mediated by function words

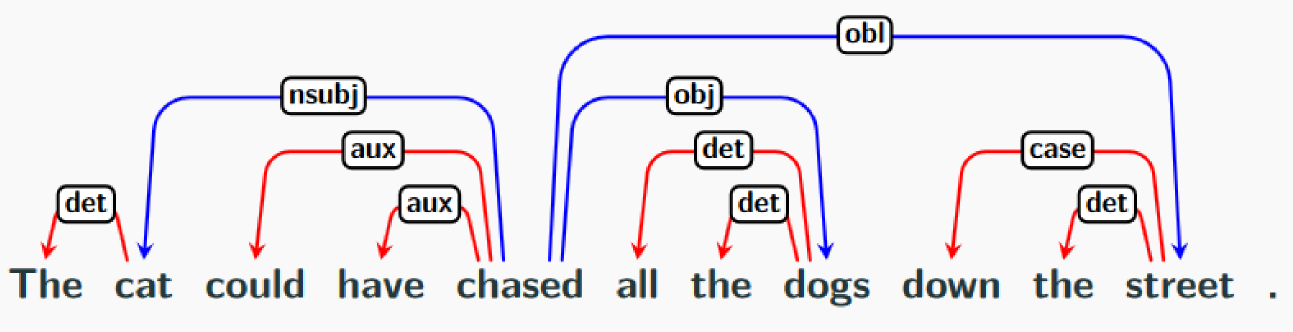


Detecting Syntactic Dependencies III

criterion of *maximal parallelism* between languages

… Universal Dependencies

- the upper levels of UD trees should be as similar as possible across languages
  - dependency relations hold primarily between content words (rather than being indirect relations mediated by function words
  - function words attach as direct dependents of the most closely related content word

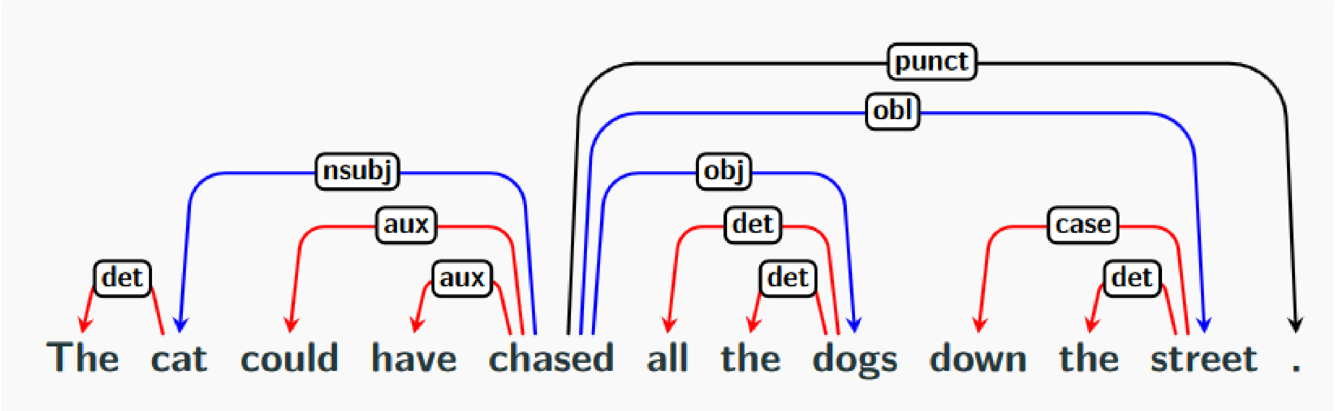


Detecting Syntactic Dependencies III

criterion of *maximal parallelism* between languages

… Universal Dependencies

- the upper levels of UD trees should be as similar as possible across languages
  - dependency relations hold primarily between content words (rather than being indirect relations mediated by function words
  - function words attach as direct dependents of the most closely related content word
  - punctuation attach to head of phrase or clause



Outline of the lecture

- Introduction: dependency grammar in a nutshell
- Tree-based structures informally
    - phrase structure / constituency trees
    - dependency trees
- How to detect a dependency relation?
- A bit of math …
- Problem with free word order

A bit of math – tree in the graph theory

tree (graph theory):
definition:

- finite graph ⟨N, E⟩, N ~ nodes/vertices, E ~ 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

A bit of math – tree in the graph theory

tree (graph theory):

definition:

- finite graph $\langle N, E\rangle$, N ~ nodes/vertices, E ~ 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 $[n_1,n_2]$)

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 ~ the ***root***

Tree as a data structure

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:
  children of each node have a specific order

Tree as a data structure

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:
  children of each node have a specific order

- "tree-ordering" D (dominance)
  partial ordering on nodes
  $u \leq_D v \iff_{\text{def}}$ the unique path from the root to v passes through u
  (weak ordering ~ reflexive, antisymmetric, transitive)
- "linear ordering" P (precedence)
  (partial) ordering on nodes
  $u <_P v$ … (strong ordering ~ antireflexive, asymmetric, transitive)

Phrase structure tree (definition, part 1)

$T = \langle N, D, Q, P, L \rangle$

⟨N, D⟩ … ***rooted tree, directed***

Q ... lexical and grammatical categories

L … labeling function N → Q

D … oriented edges (branches)

~ relation on lexical and grammatical categories

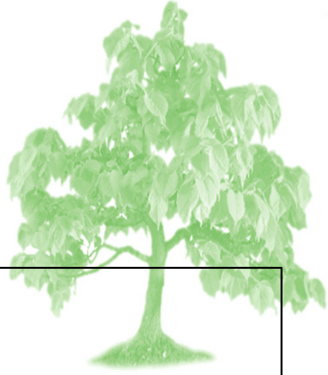
dominance relation

+

P ... relation on N ~ (partial strong linear ordering)

relation of *precedence*

Phrase structure tree (definition, part 2)



T = ⟨ *N, D, Q, P, L* ⟩

⟨N, D⟩ … ***rooted tree, directed***

Q ... lexical and grammatical categories

L … labeling function N → Q

D … oriented edges (branches)

~ relation on lexical and grammatical categories

dominance relation

+

P ... relation on N ~ (partial strong linear ordering)

relation of *precedence*

+

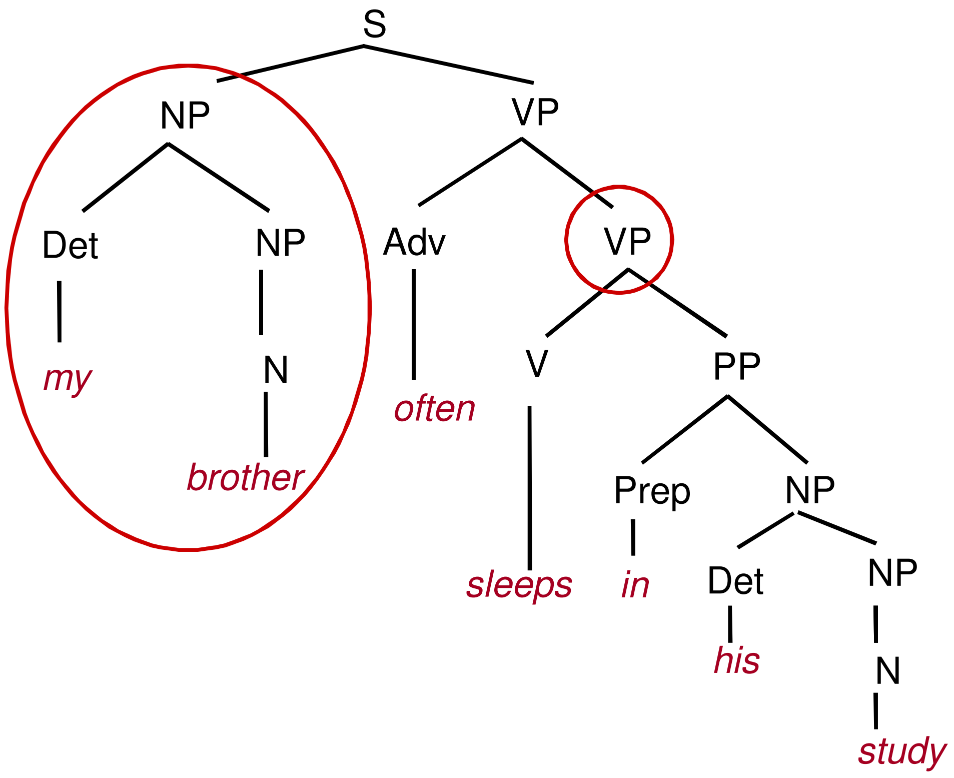
Relating dominance and precedence relations:

- *exclusivity* condition for D and P relations
- *'nontangling'* condition

Phrase structure tree (relation P)

- *exclusivity* condition for D and P relations

$\forall\ x,y \in N$ holds: $([x,y] \in P \vee [y,x] \in P) \Leftrightarrow ([x,y] \notin D \ \&\ [y,x] \notin D)$



Phrase structure tree (relation P)

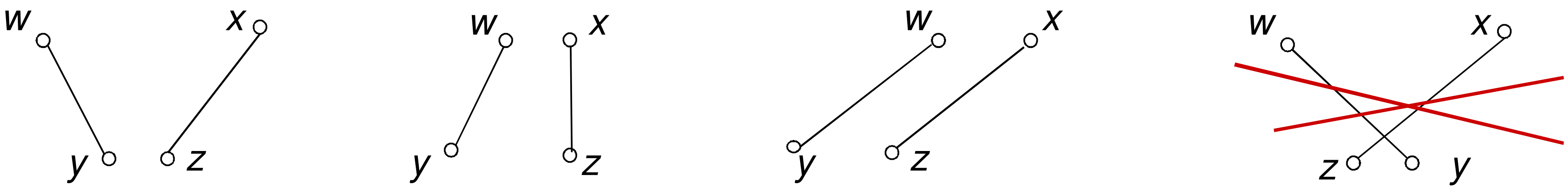
- *exclusivity* condition for D and P relations

  $\forall\ x,y \in N$ holds: $(\ [x,y] \in \mathrm{P} \vee [y,x] \in \mathrm{P}\ ) \Leftrightarrow (\ [x,y] \notin \mathrm{D}\ \&\ [y,x] \notin \mathrm{D})$

- *‘nontangling’* condition

  $\forall\ w,x,y,z \in N$ holds: $(\ [w,x] \in \mathrm{P}\ \&\ [w,y] \in \mathrm{D}\ \&\ [x,z] \in \mathrm{D}\ )$

  $\Rightarrow (\ [y,z] \in \mathrm{P}\ )$



Phrase structure tree (relation P)

- *exclusivity* condition for D and P relations

  $\forall\ x,y \in N$ holds: $(\ [x,y] \in \mathrm{P} \vee [y,x] \in \mathrm{P}\ ) \Leftrightarrow (\ [x,y] \notin \mathrm{D}\ \&\ [y,x] \notin \mathrm{D})$

- *‘nontangling’* condition

  $\forall\ w,x,y,z \in N$ holds: $(\ [w,x] \in \mathrm{P}\ \&\ [w,y] \in \mathrm{D}\ \&\ [x,z] \in \mathrm{D}\ )$

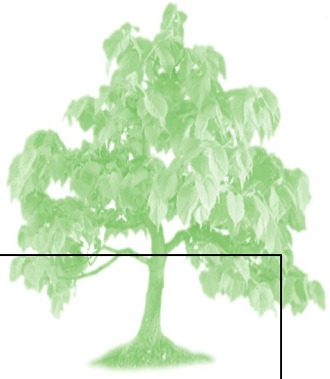
  $\Rightarrow (\ [y,z] \in \mathrm{P}\ )$

⇨

$T = \langle N,D,Q, P,L \rangle$ phrase structure tree

- $\forall\ \ x,y \in \mathrm{N}$ siblings $\Rightarrow [x,y\,] \in \mathrm{P}$
- the set of its leaves is totally ordered by P

Dependency tree (definition)



$T = \langle N, D, Q, WO, L \rangle$

$\langle N, D \rangle$ … ***rooted tree, directed***

Q ... lexical and grammatical categories

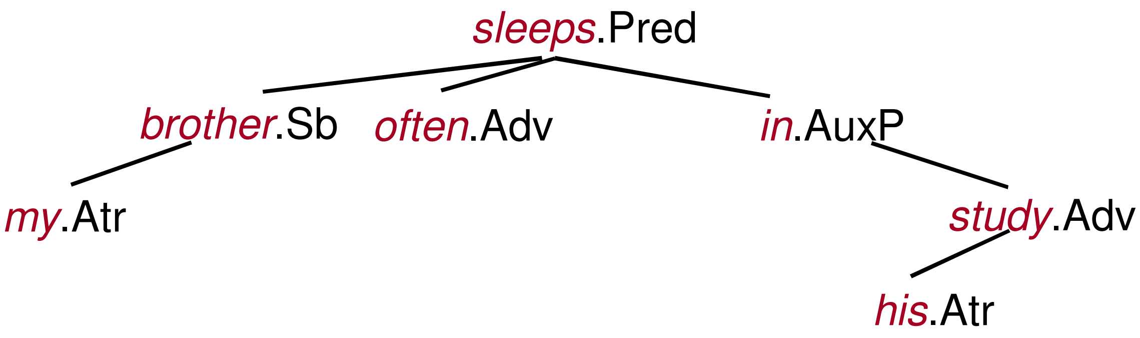
L … labeling function $N \rightarrow Q^+$

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

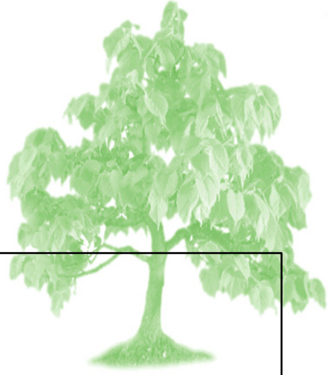
‘dependency’ relation

WO ...relation on N ~ (strong total ordering on N) …

word order



Subtree vs. catena



$T = \langle N, D, Q, WO, L \rangle$

$\langle$N, D$\rangle$ … ***rooted tree, directed edges***

Subtree

- Purely set-wise: $N_1 \subseteq N$, the other items (D etc.) reduced accordingly
- (Complete) subtree: Take all nodes from N that are dominated by a given node u: $N_1 = \{ v \mid u \leq_D v \}$

Catena (Osborne and Groß 2016)

- **Connected** subgraph: $N_1 \subseteq N$, the other items (D etc.) reduced accordingly. There is one (and only one) node u that is not immediately dominated by any other node in N_1.

Outline of the lecture

- Introduction: dependency grammar in a nutshell
- Tree-based structures informally
  - phrase structure / constituency trees
  - dependency trees
- How to detect a dependency relation?
- A bit of math …
- Problem with free word order

Problem with free word order

free word order:

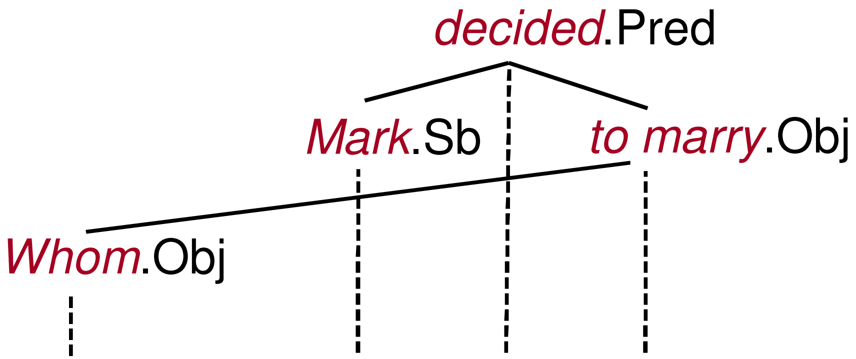
- freedom of word order of dependents within a continuous ‘head domain’ (i.e., substring of head + its dependents)

Problem with Free Word Order

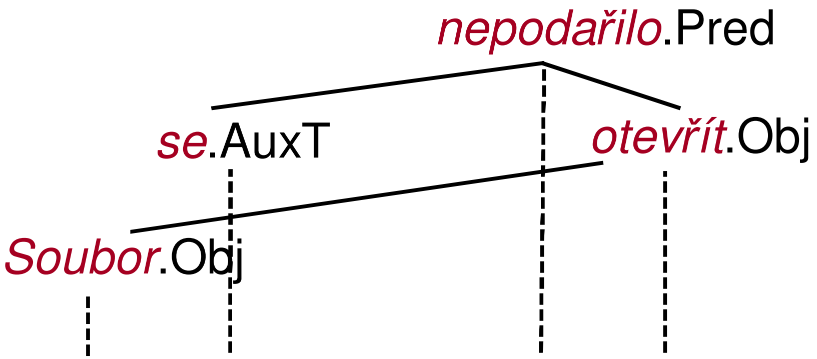
free word order:

- freedom of word order of dependents within a continuous 'head domain' (i.e., substring of head + its dependents)
- relaxation of continuity of a head domain

Whom did Mark decide to marry?



Soubor se mi nepodařilo otevřít. (Oliva)



Problem with Free Word Order

free word order:

- freedom of word order of dependents within a continuous 'head domain' (i.e., substring of head + its dependents)
- relaxation of continuity of a head domain

German:
Maria hat einen Mann kennengelernt der Schmetterlinge sammelt.
Mary has a man met the butterflies collects
'Mary has met a man who collects butterflies.'

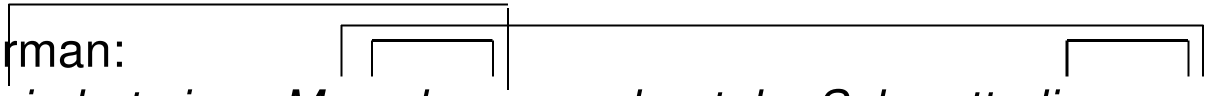
English: long-distance unbounded dependency
John, Peter thought that Sue said that Mary loves.

Czech:
Marii se Petr tu knihu rozhodl nekoupit.
to-Mary PART Peter that book decided not-buy
'Peter decided not to buy that book to Mary.'

Problem with Free Word Order

free word order:

- freedom of word order of dependents within a continuous 'head domain' (i.e., substring of head + its dependents)
- relaxation of continuity of a head domain



German:
Maria hat einen Mann kennengelernt der Schmetterlinge sammelt.
Mary has a man met the butteries collects
'Mary has met a man who collects butteries.'

English: long-distance unbounded dependency
John, Peter thought that Sue said that Mary loves.

Czech*:*
Marii se Petr tu knihu rozhodl nekoupit.
to-Mary PART Peter that book decided not-buy
'Peter decided not to buy that book to Mary.'

References

- Osborne, T. (2019) *A Dependency Grammar of English*. John Benjamins Publishing Company, Amsterdam/Philadelphia
- Mel’čuk, I. (1988) *Dependency Syntax: Theory and Practice.* State University of New York Press, Albany
- Sgall, P., Hajičová, E., Panevová, J. (1986) *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht/Academia, Prague
- Hajičová, E., Panevová, J., Sgall, P. (2002) *Úvod do teoretické a počítačové lingvistiky*, sv. I. Karolinum, Praha
- Osborne, T., Groß, T. (2016) The do-so-diagnostic: Against finite VPs and for flat non-finite VPs. In: *Folia linguistica* 50, 1, 97–35
- Partee, B. H.; ter Meulen, A.; Wall, R. E. (1990) *Mathematical Methods in Linguistics*. Kluwer Academic Publishers
- Petkevič, V. (1995) A New Formal Specification of Underlying Structure. *Theoretical Linguistics*, vol. 21, No.1
- Štěpánek, J. (2006*) Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu*. PhD Thesis, MFF UK
- Universal Dependencies https://universaldependencies.org
- Prague Dependency Treebank http://ufal.mff.cuni.cz/pdt3.5