IB047

Syntaktické značkování korpusů

Pavel Rychlý

pary@fi.muni.cz

28. března 2017

Morfologické značkování

- každý token značka
- několik desítek až tisíc značek (obsahující gramatické kategorie)
- Universal Tagset (Google)
  12 značek – pouze slovní druhy
- jeden sloupec ve vertkálním tvaru

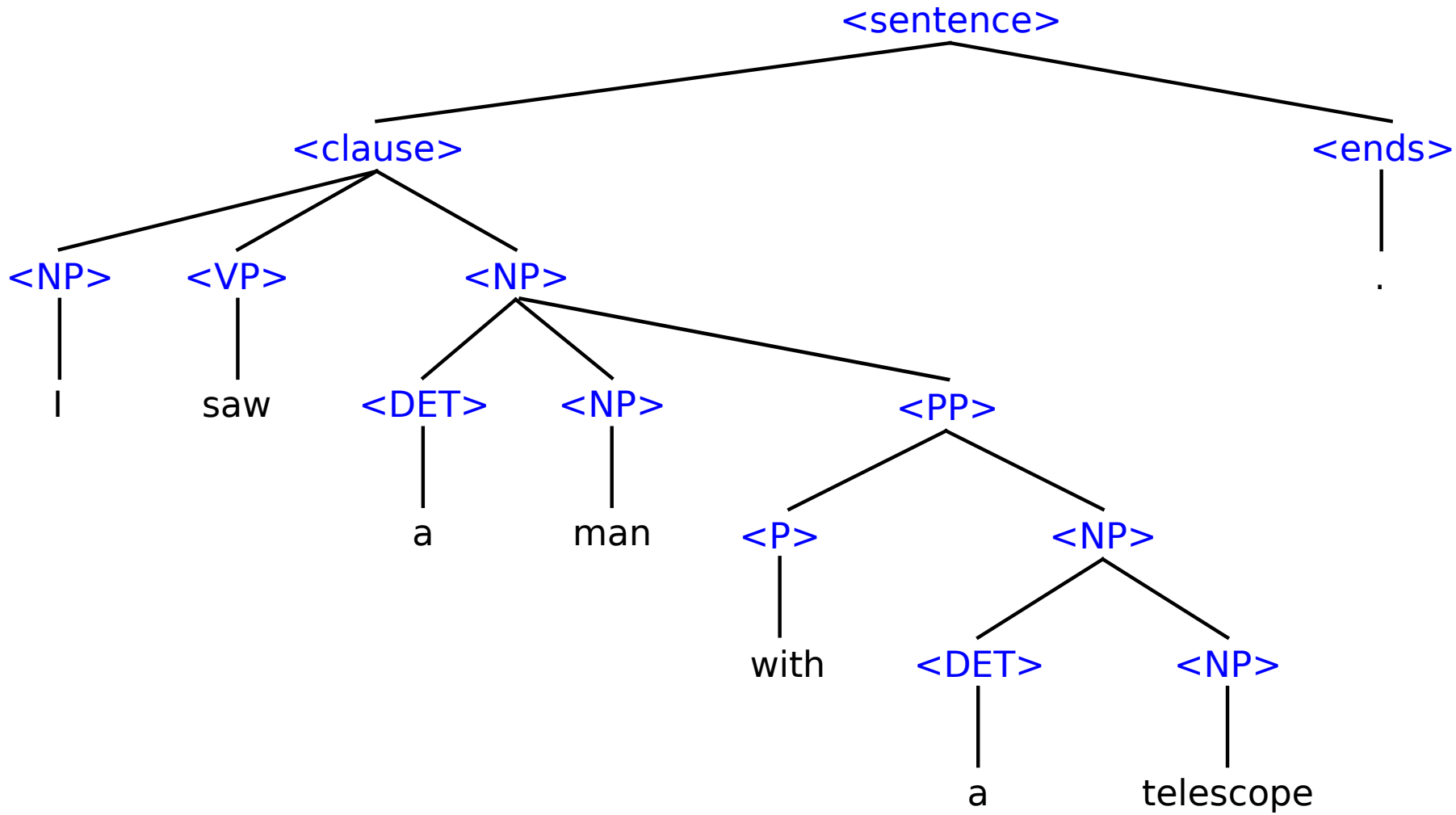
Universal Dependencies

- nástupce Universal Tagset
- verze 2.0 (12/2016, 3/2017): 70 korpusů, 50 jazyků
- 14 značek – slovní druhy (+ 3 ostatní)
- open class
  ADJ: adjective, ADV: adverb, INTJ: interjection, NOUN: noun, PROPN: proper noun, VERB: verb
- closed class
  ADP: adposition, AUX: auxiliary, CCONJ: coordinating conjunction, DET: determiner, NUM: numeral, PART: particle, PRON: pronoun, SCONJ: subordinating conjunction
- ostatní
  PUNCT: punctuation, SYM: symbol, X: other
- 21 features
  PronType, Gender, Animacy, Number, Tense, Abbr, ...

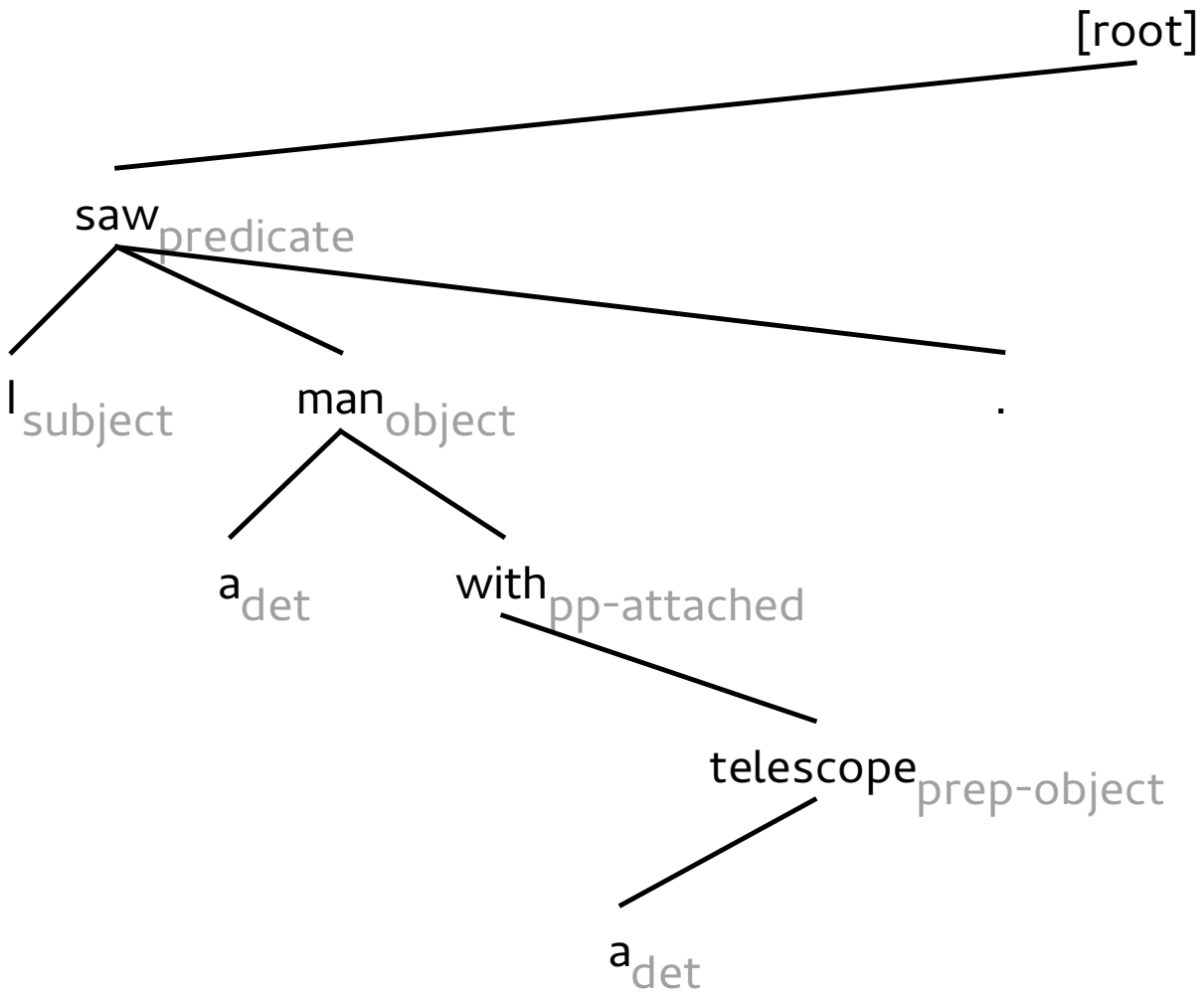
Přístupy k syntaxi

- pro každou větu vytvoříme strom zachycující vztahy mezi slovy a/nebo skupinami slov
- frázový (složkový)
  postupně ze slov vytváříme skupiny
- závislostní
  určujeme závislosti mezi jednotlivými slovy

Phrase structure formalism – example



Dependency formalism – example



Dependency vs. phrase-structure

Non-projectivity

- disconnected phrases
- not natural in the phrase structure notation
- 20% of Czech sentences are reported to contain a non-projective dependency

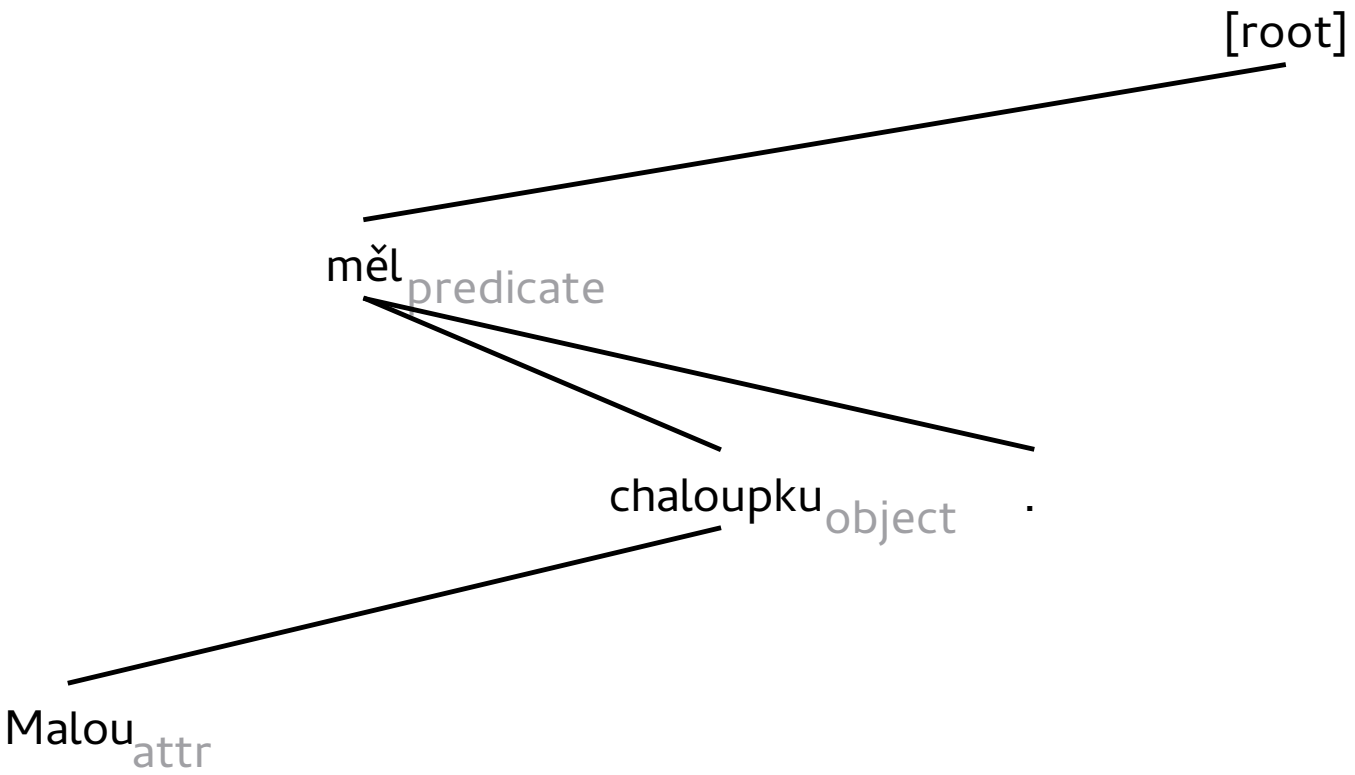
Phrase structure – more fine-grained analysis

- (new (queen of beauty))
- (new generation)(of fighters)

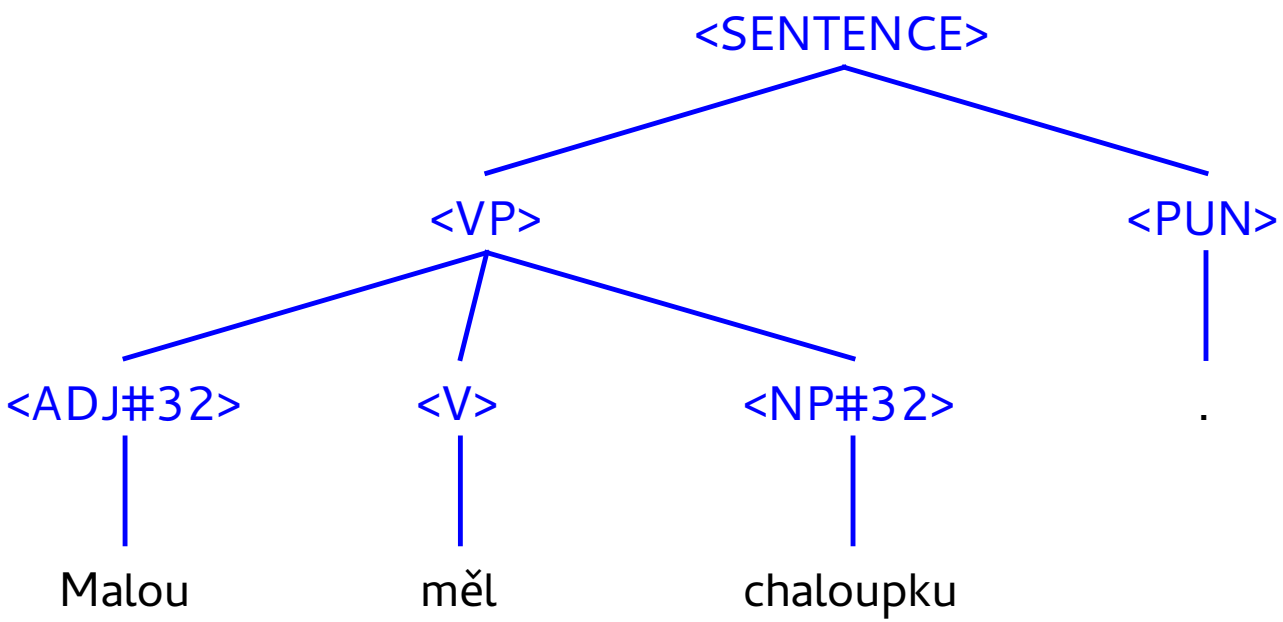
Coordinations and other “flat” phenomena

- not natural in the dependency notation
- problem for dependency analysis

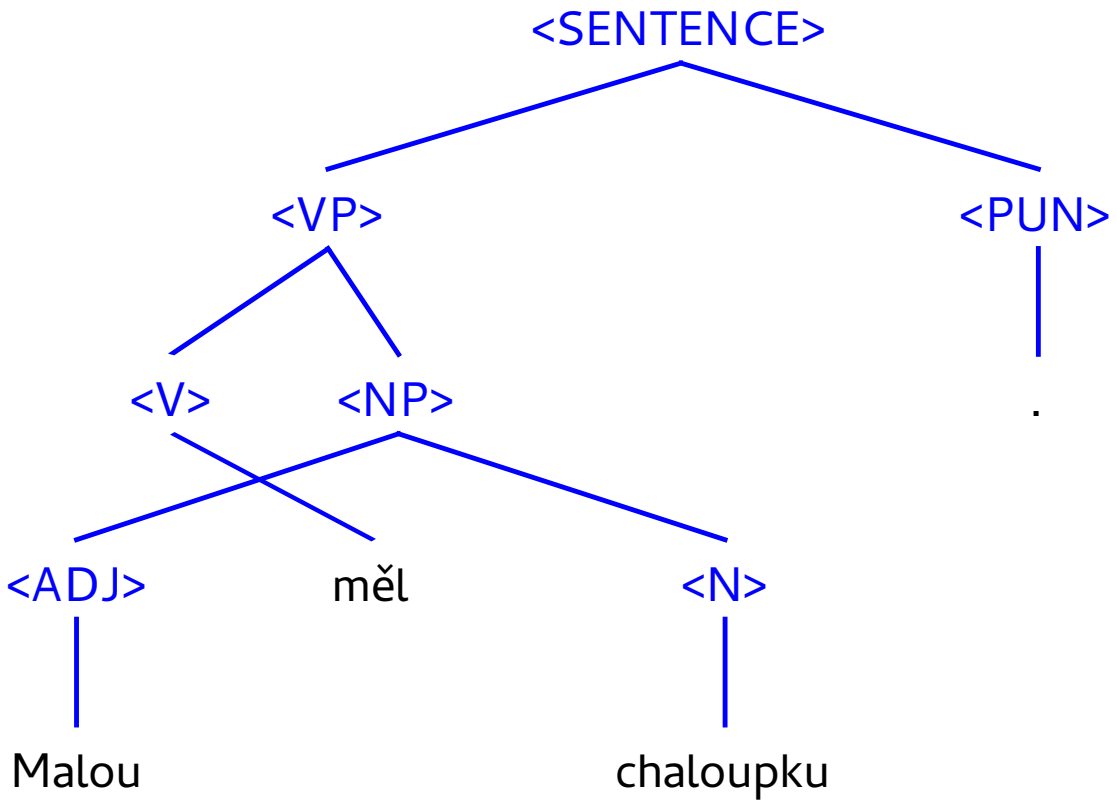
Non-projectivity – example



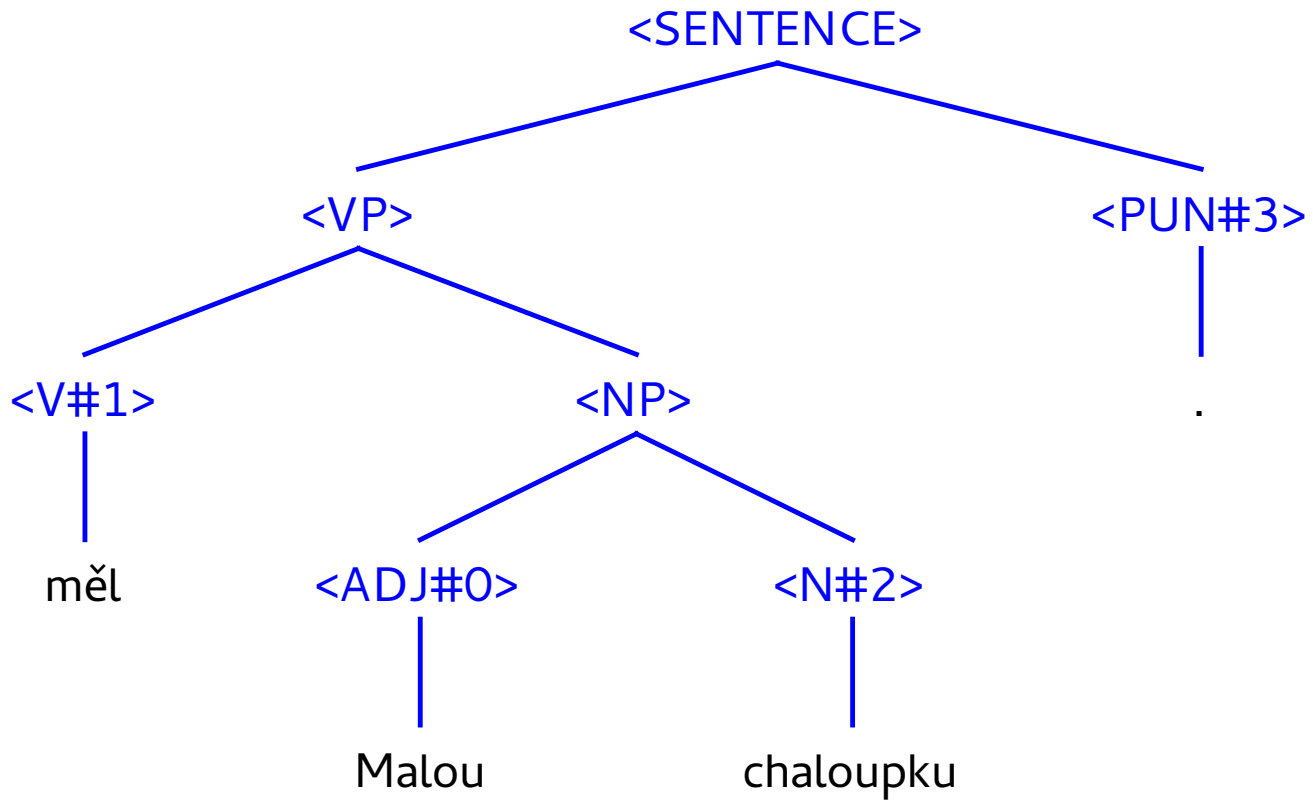
Non-projectivity in phrase structure formalism



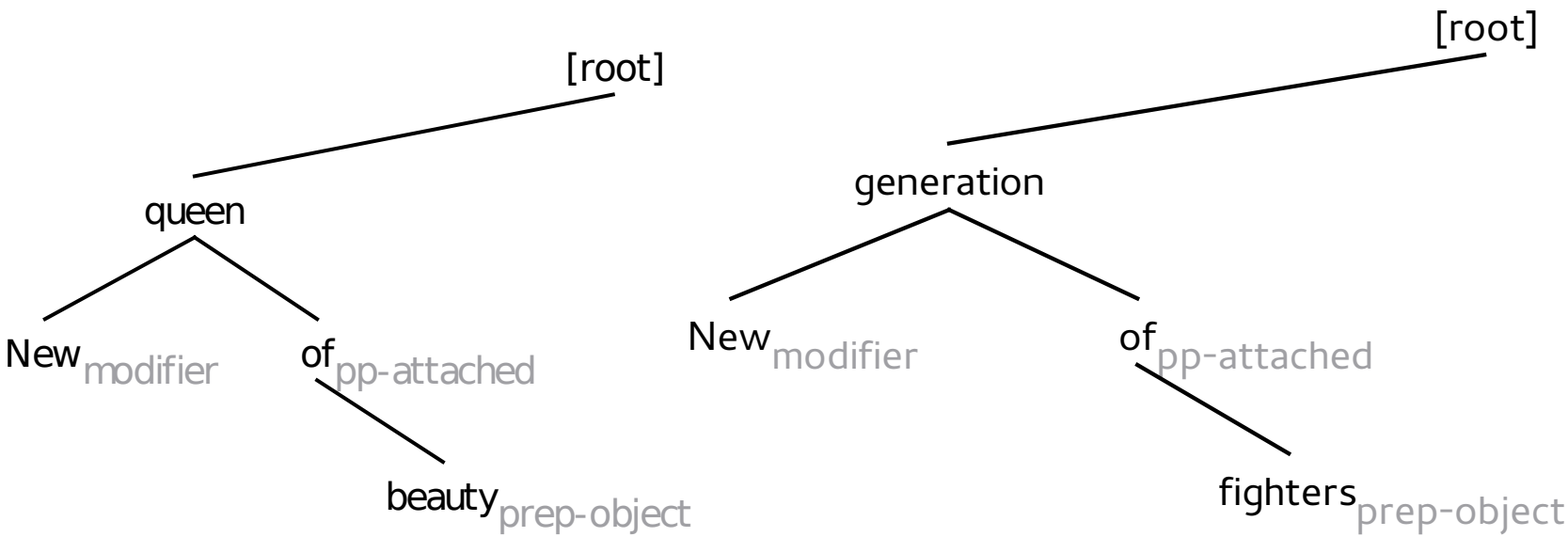
Non-projectivity in phrase structure formalism



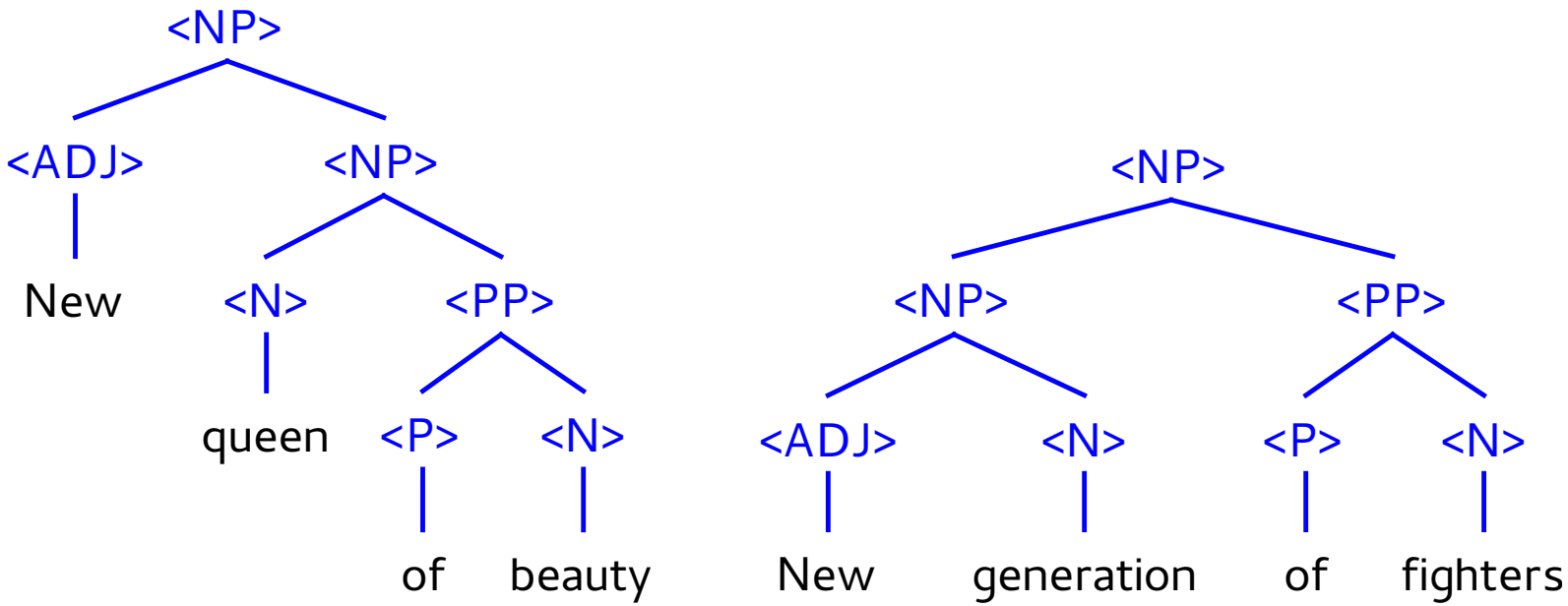
Non-projectivity in phrase structure formalism



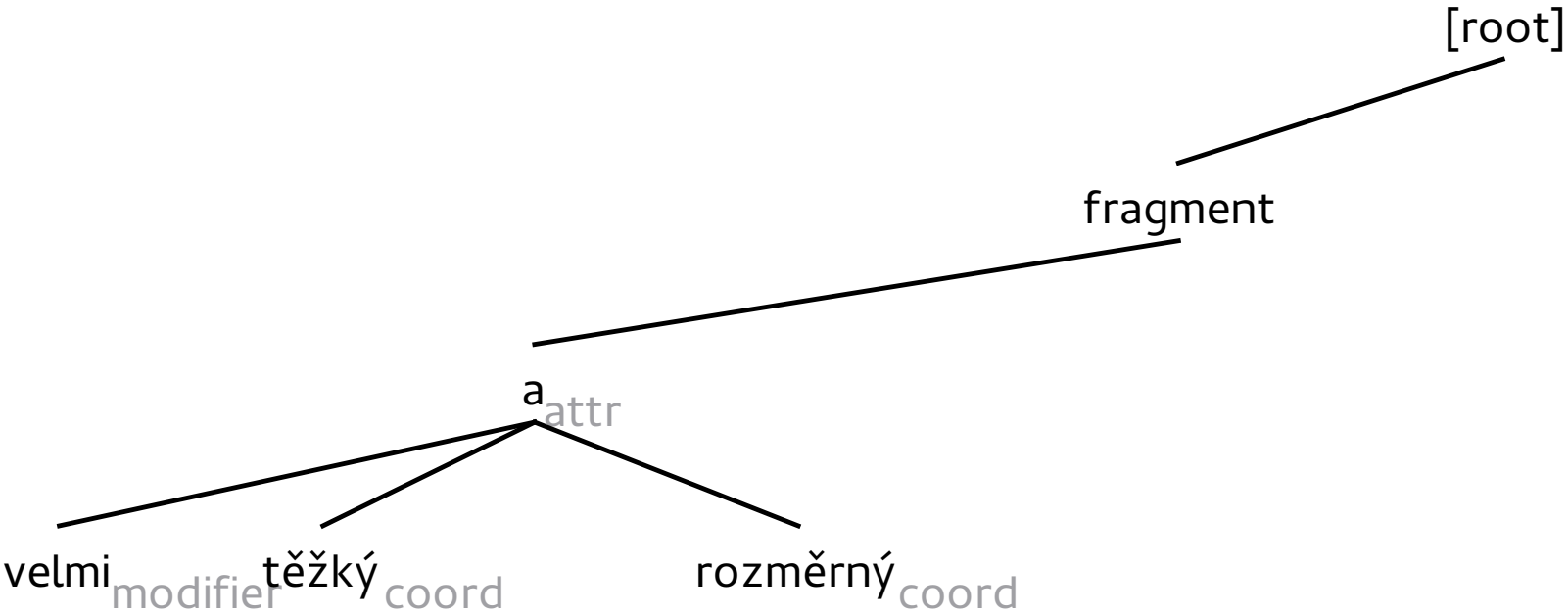
Phrase structure expressivity



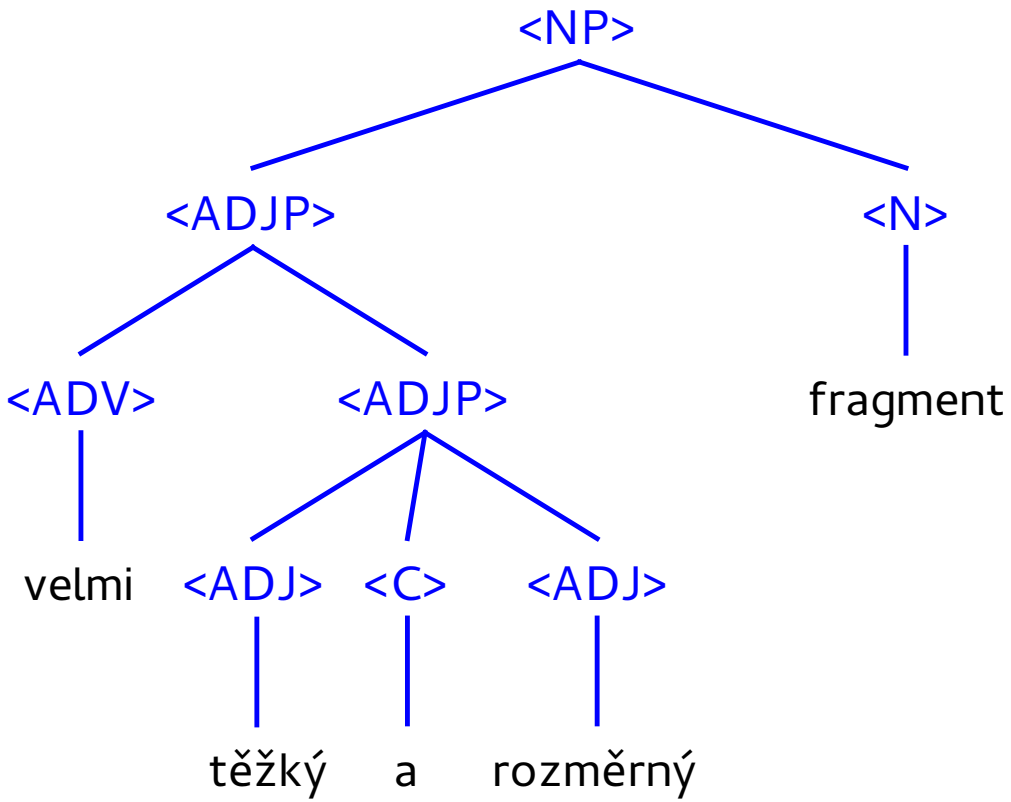
Phrase structure expressivity



Coordinations – dependency structure



Coordinations – phrase structure



Záznam syntaxe v korpusu

- složkový systém:
  - fráze jako struktury – `<phr>, </phr>`
  - typy jako atributy
- závislostní systém:
  - očíslování tokenů, odkazy na tokeny
  - typy relací v dalším atributu