IB047

Syntaktické značkování korpusů

Pavel Rychlý

pary@fi.muni.cz

17. března 2014

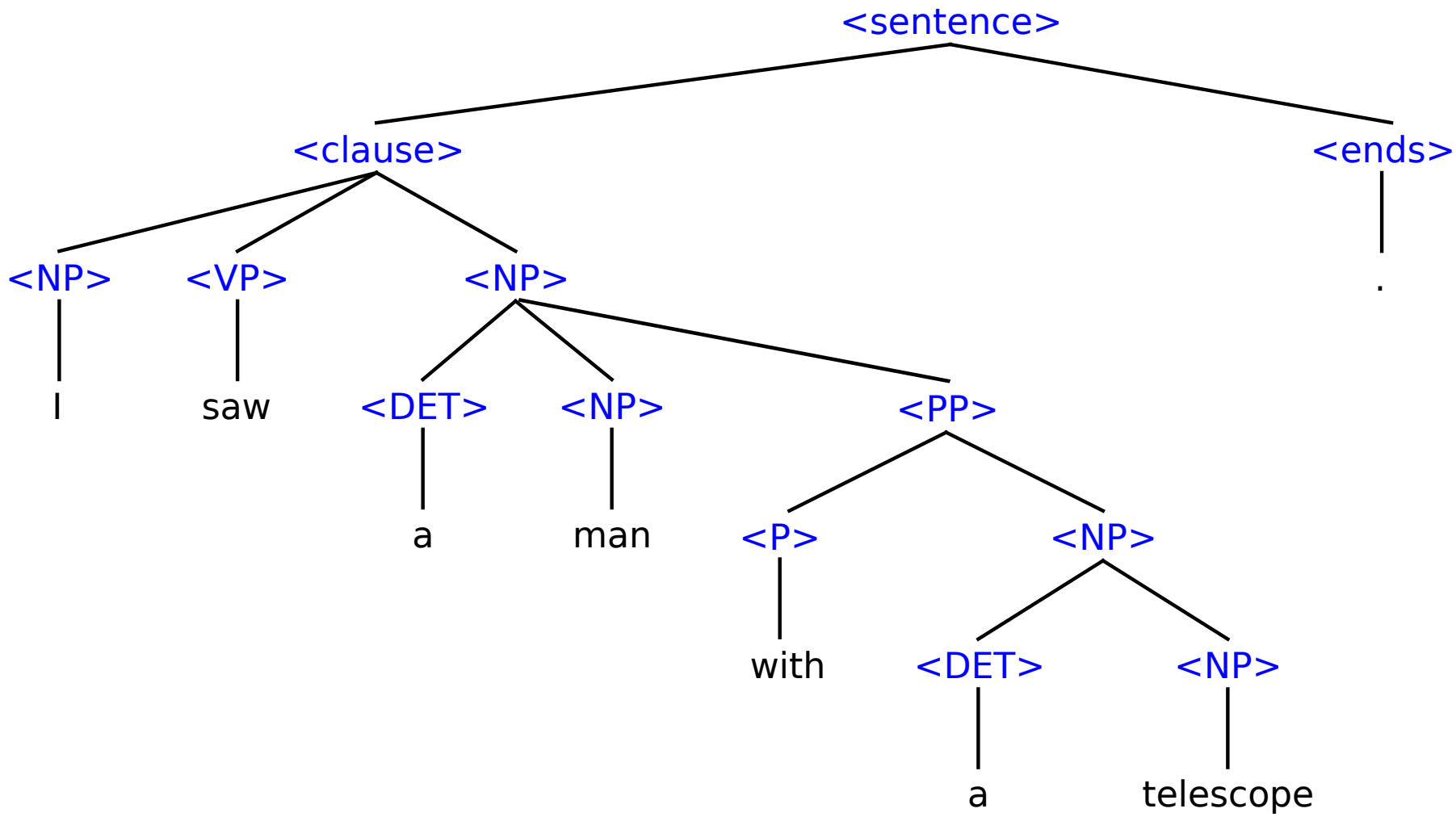
Morfologické značkování

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

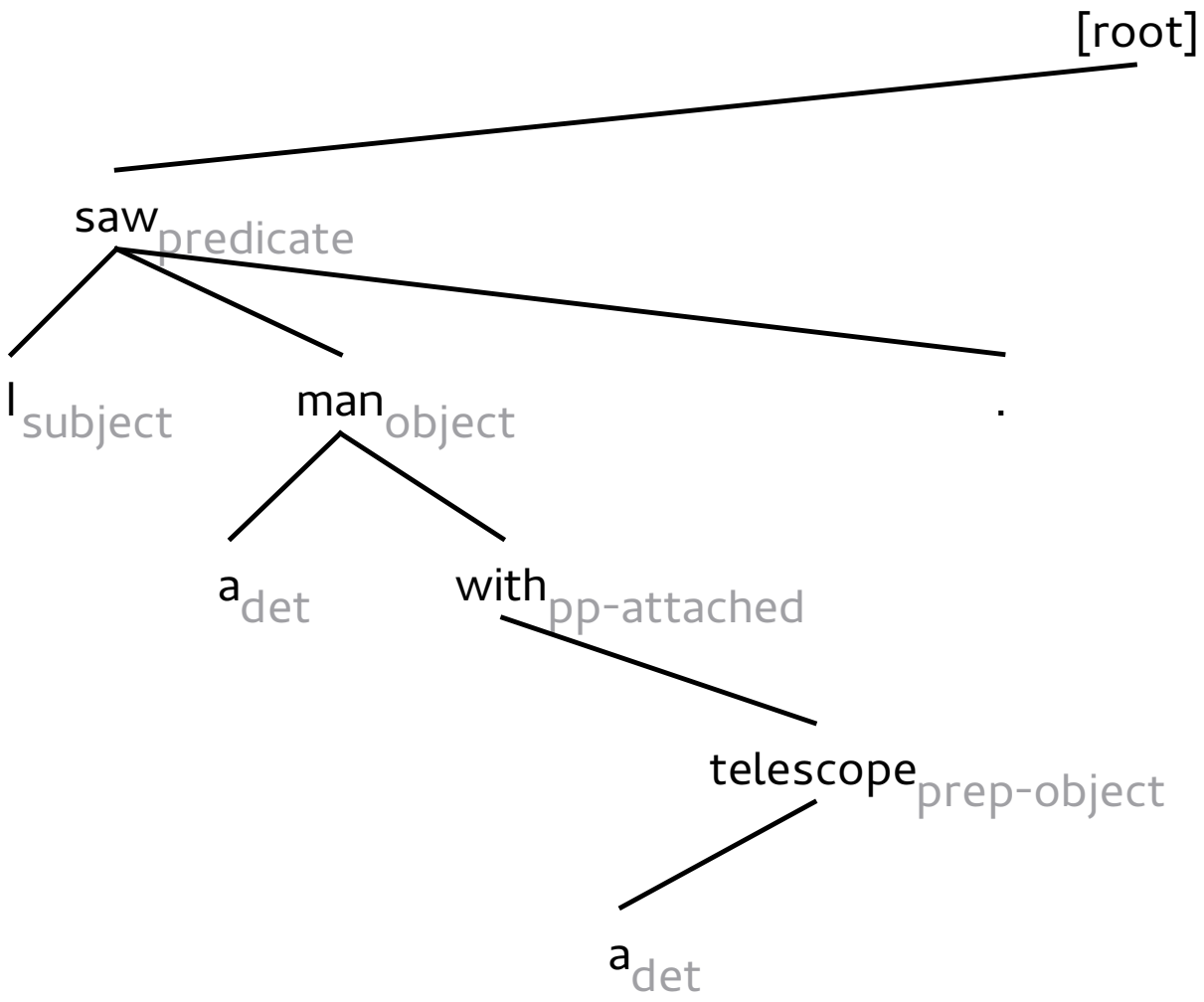
Přístupy k syntaxi

- 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



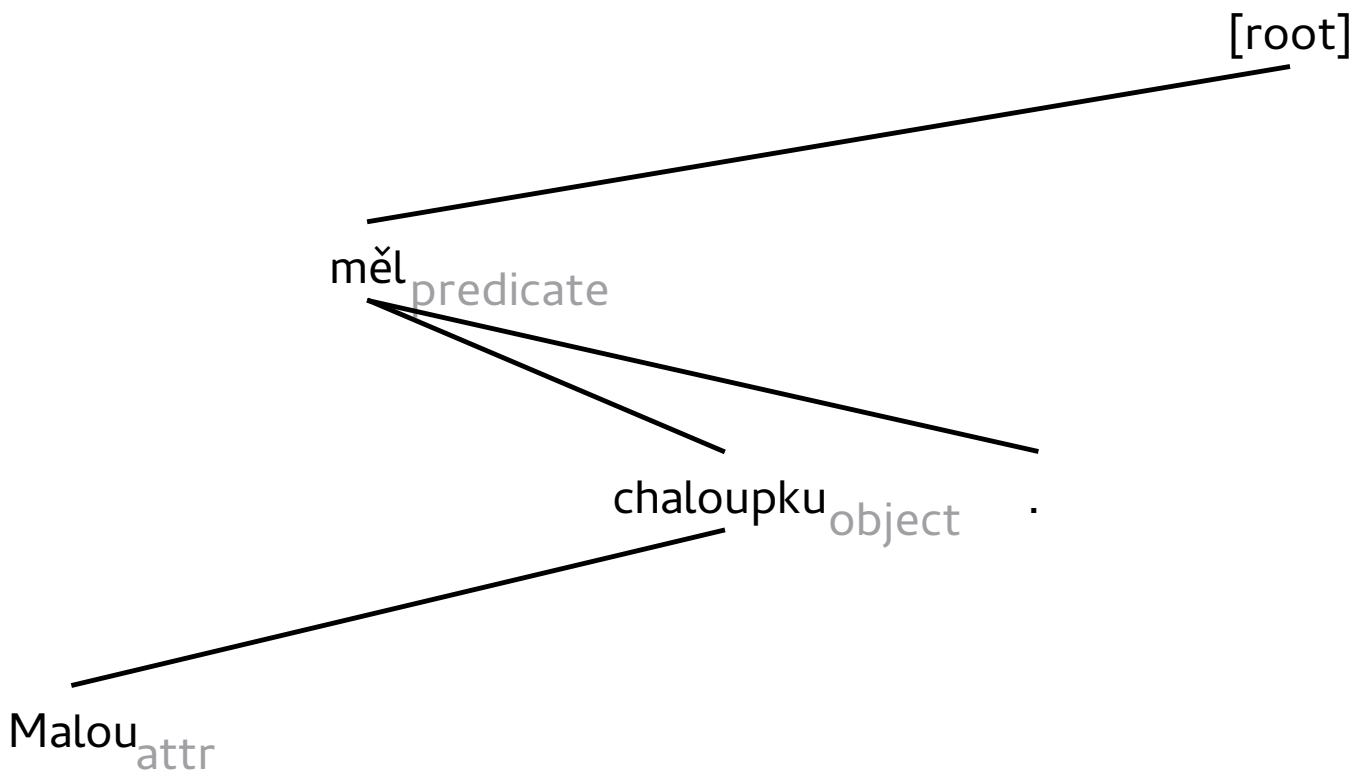
Dependency formalism – example



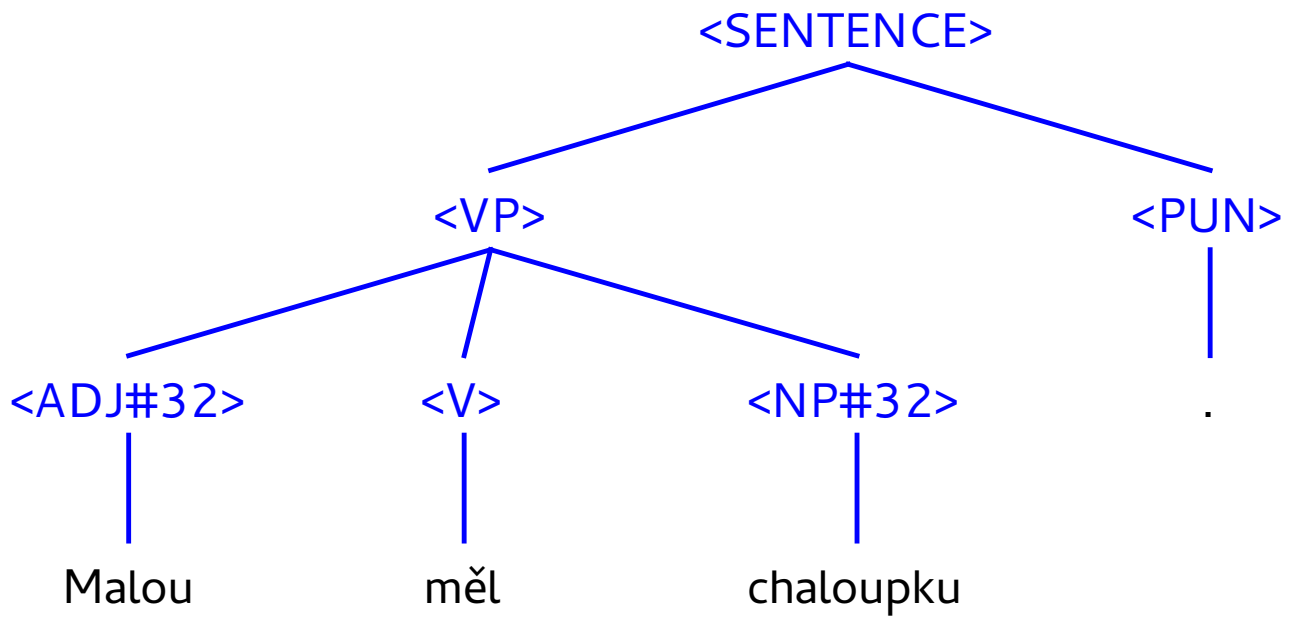
Dependency vs. phrase-structure

- 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

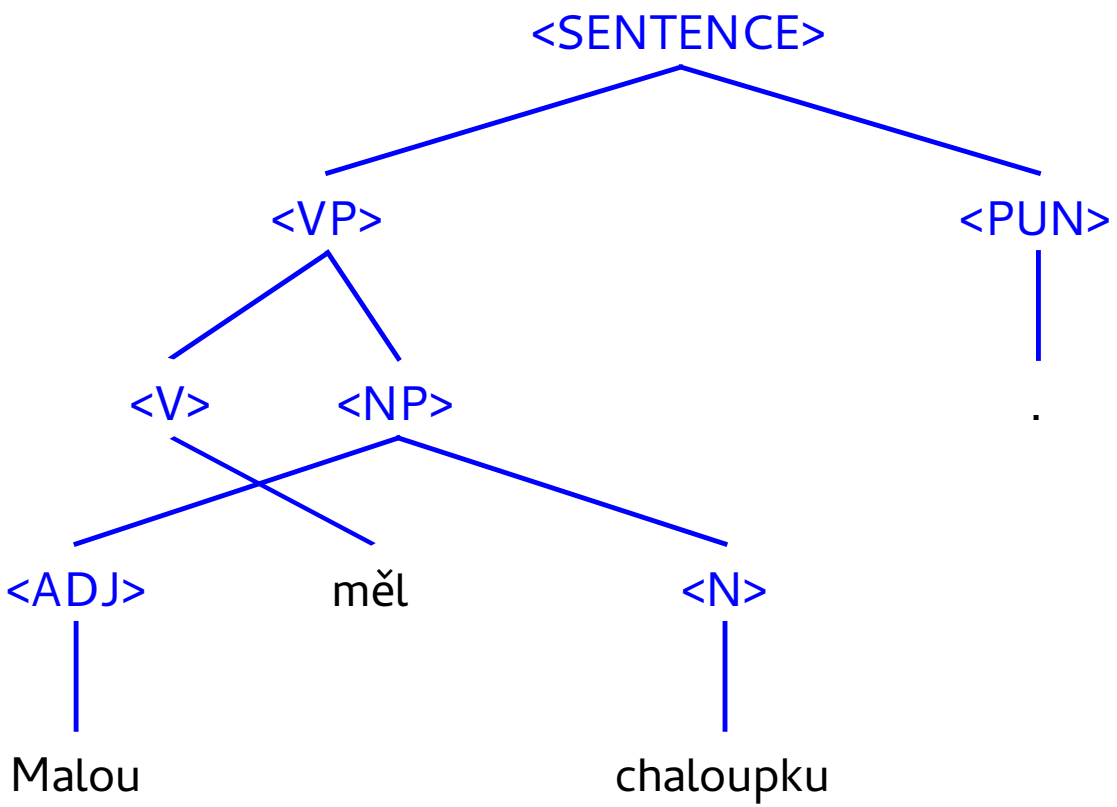
Non-projectivity – example



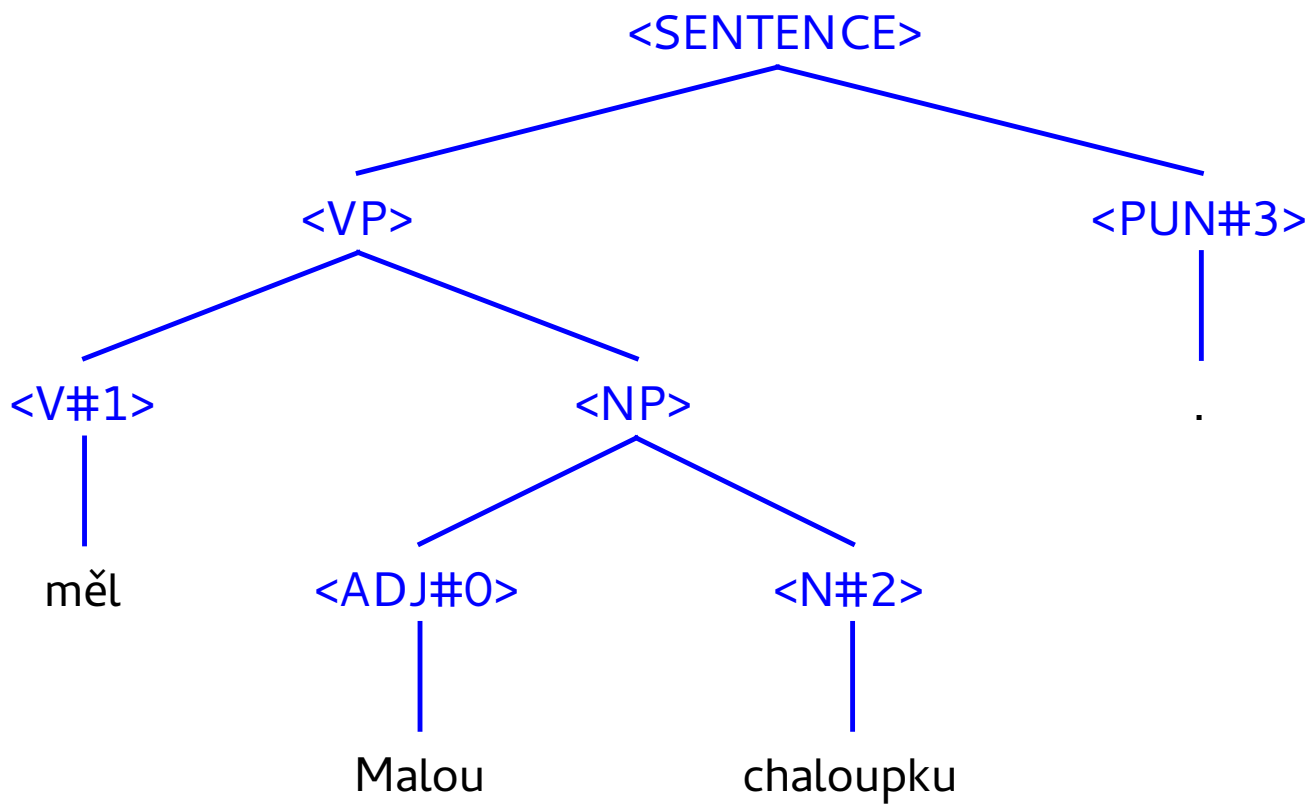
Non-projectivity in phrase structure formalism



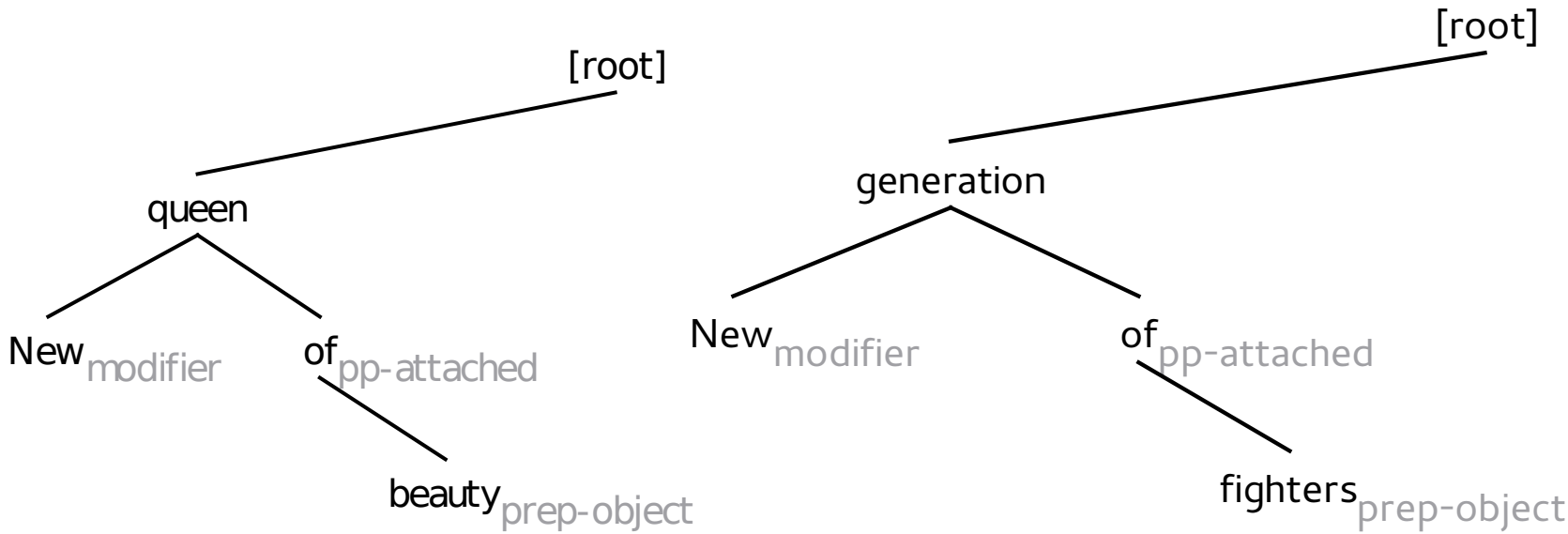
Non-projectivity in phrase structure formalism



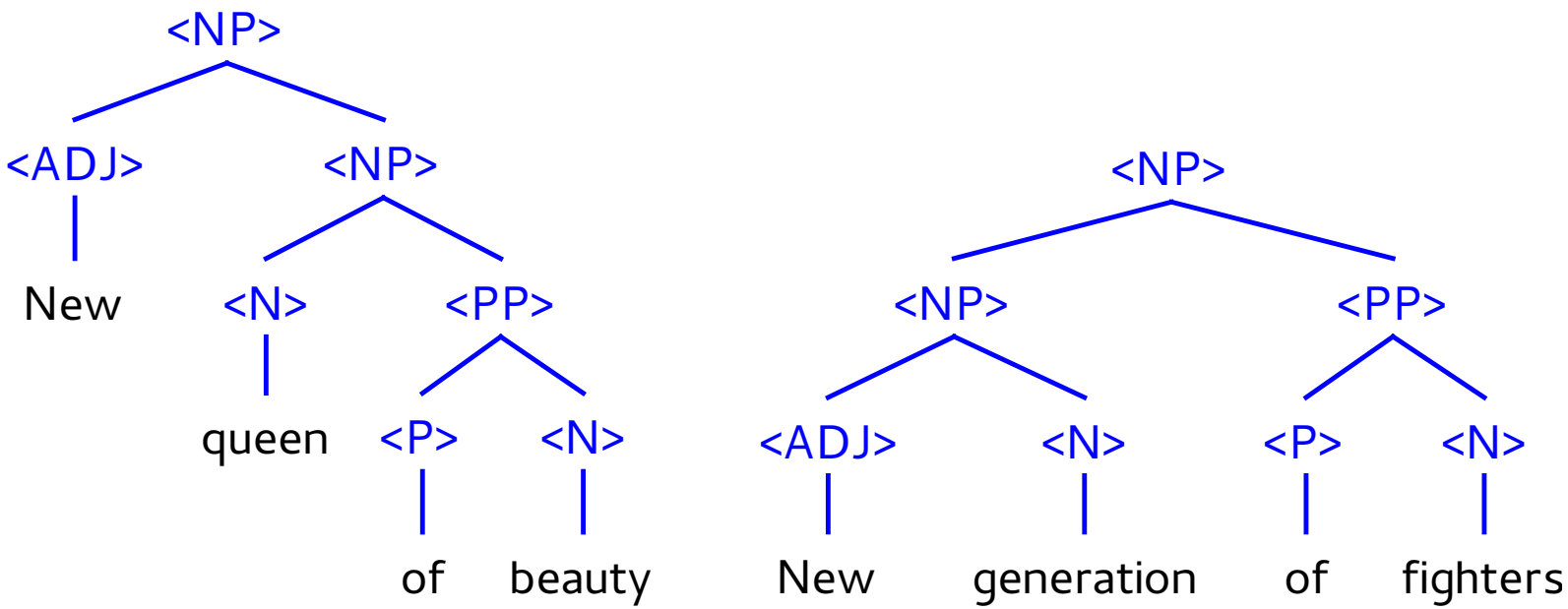
Non-projectivity in phrase structure formalism



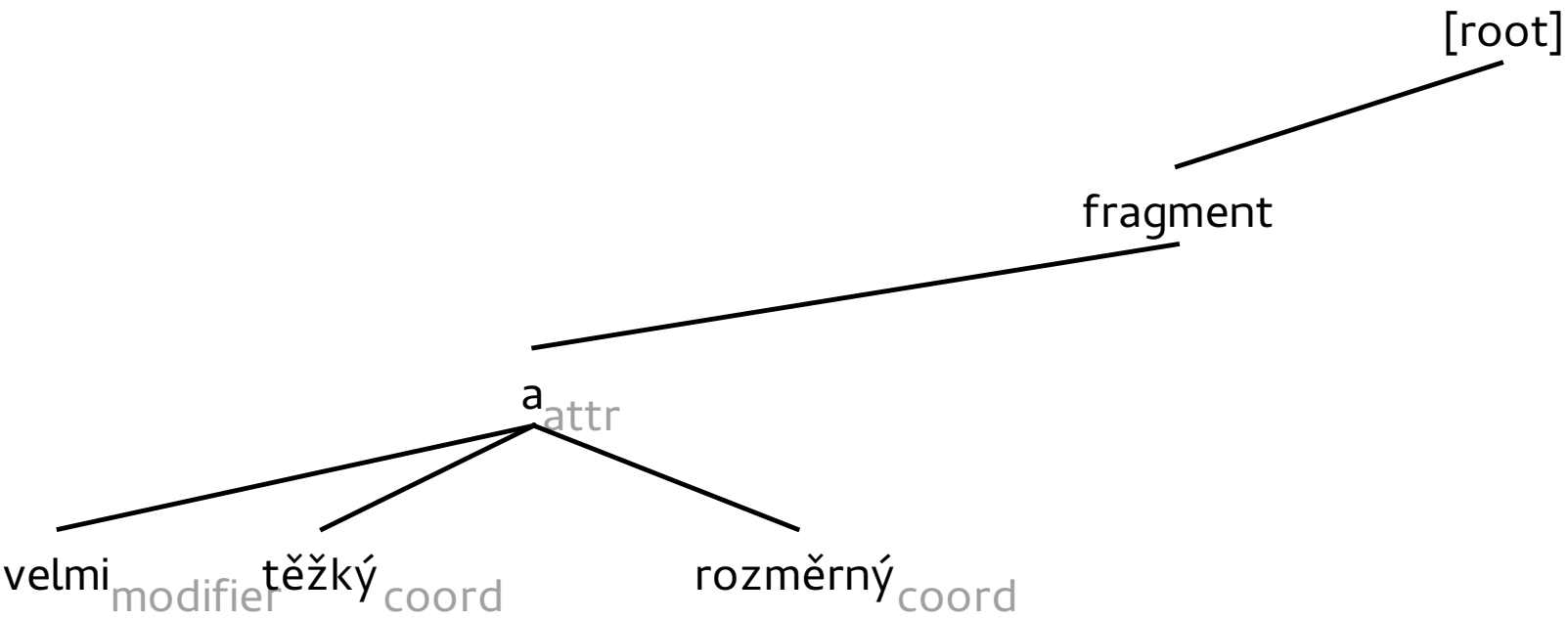
Phrase structure expressivity



Phrase structure expressivity



Coordinations – dependency structure



Coordinations – phrase structure