Syntaktická analýza

PLIN059

Mgr. Dana Hlaváčková, Ph.D.

Syntaktická analýza

- počítačové zpracování věty
  - lineární řetězec tokenů
  - graf (vztahy větných členů) – **strom (tree)**
- rozpoznání hranice věty – segmenter (statistický, pravidlový)
  - kde věta začíná a končí (velké počáteční písmeno, interpunkce)
    - *...nechutnalo nám.*
    - *...Masarykovo nám. č. 13.*

(V SYN2015 jen 4 x náměstí, 18 x označkováno jako N)

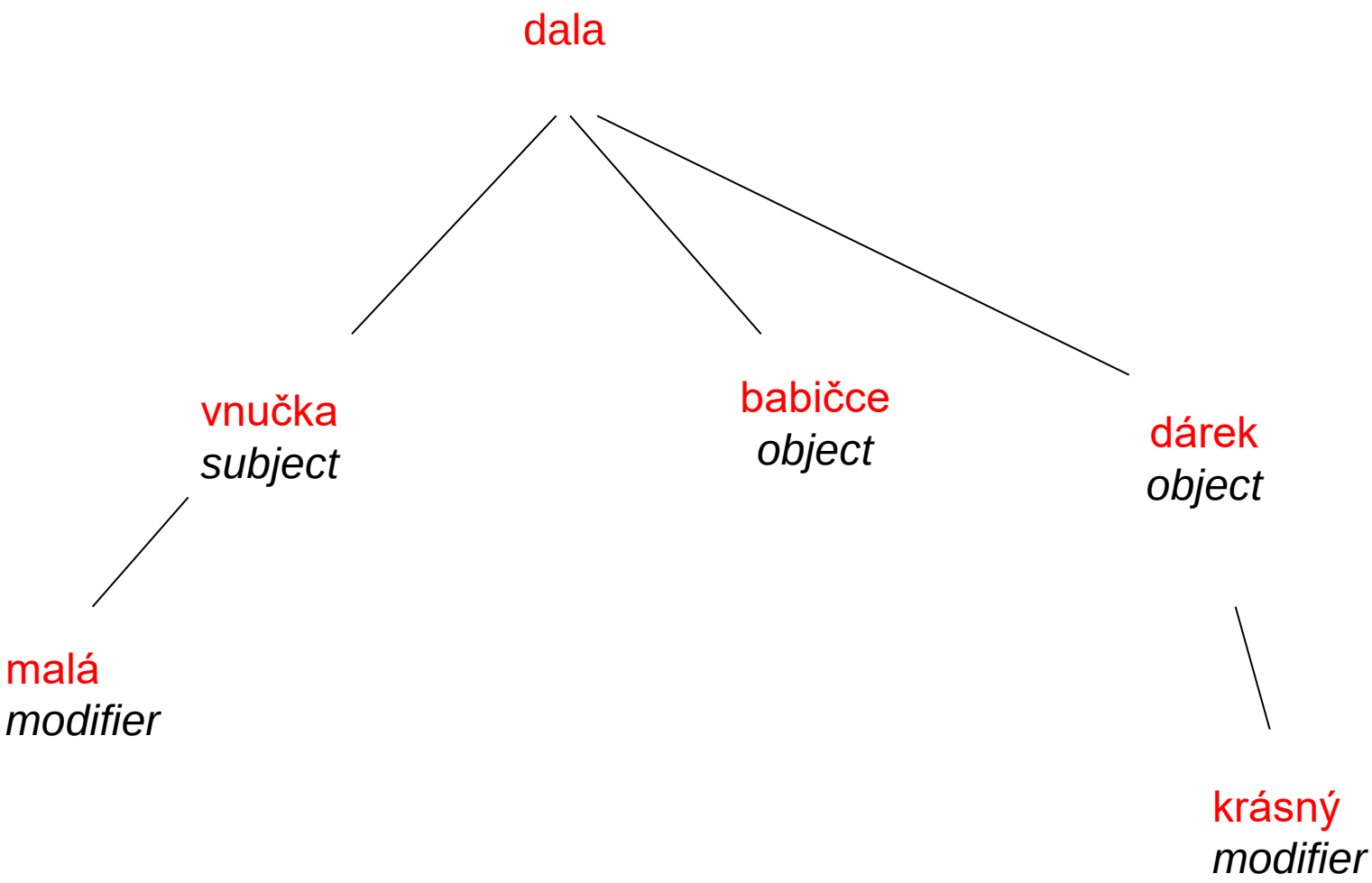
Syntaktická analýza

- předpoklad
  - **tokenizace** – tokeny (listy)
  - **morfologicky** správně označkovaný (a desambiguovaný) korpus
  - správná **segmentace** vět
- **stromy** – uzly a hrany
  - **závislostní** (řídicí a podřízené členy)
  - **složkové** (bezprostřední složky – fráze)

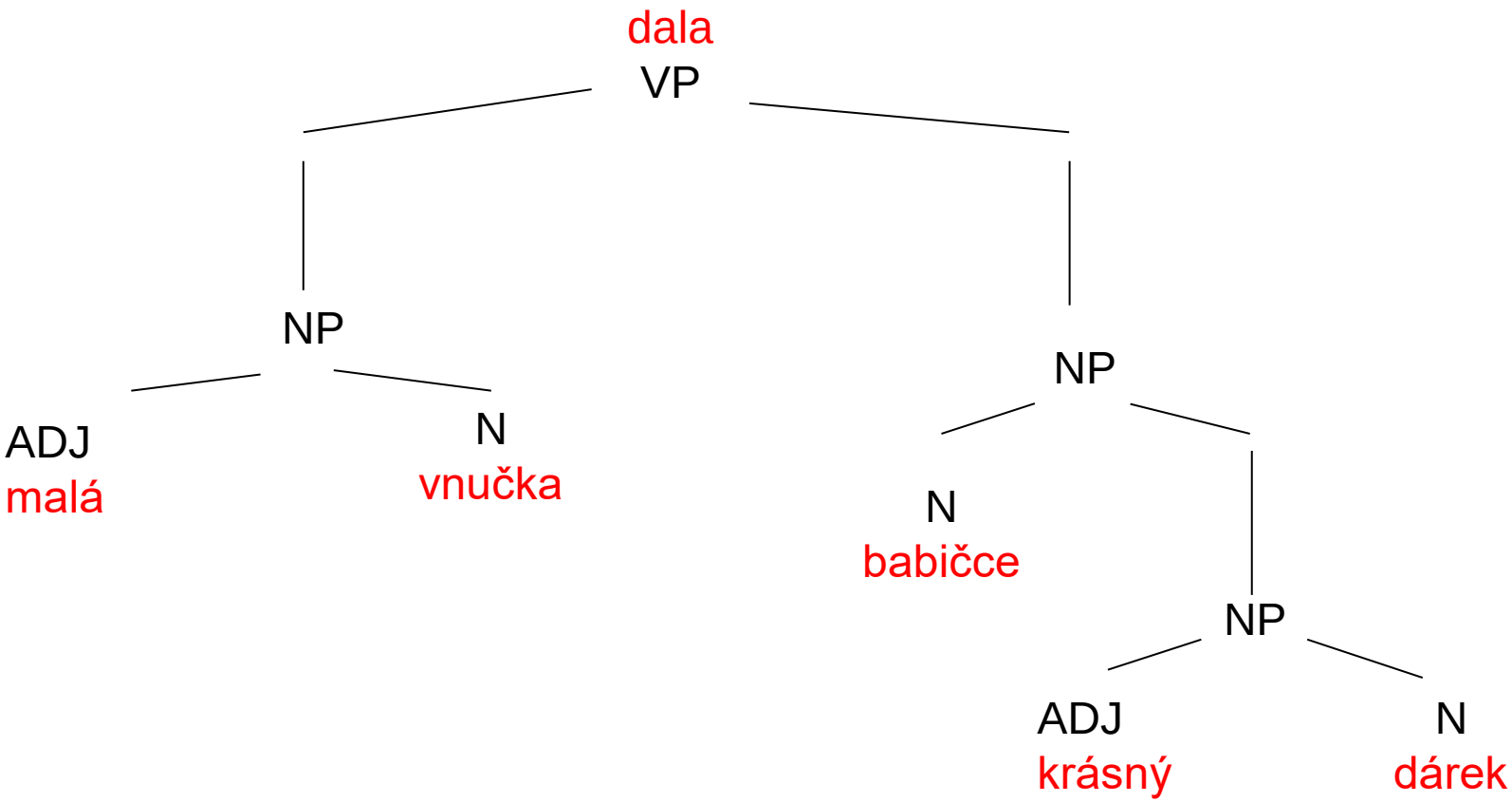
Základní pojmy

- stromy (větve, listy) – tree
- stromová banka, závislostní korpus – treebank
- syntaktická analýza – parsing
- syntaktický analyzátor – parser
- věta – sentence S, klauze – clause
- nominální fráze – NP (nominal phrase)
- verbální fráze – VP (verbal phrase)

Malá vnučka dala babičce krásný dárek. ZÁVISLOSTNÍ STROM



Malá vnučka dala babičce krásný dárek. SLOŽKOVÝ STROM



Syntaktická analýza

- **syntaktický analyzátor**
  - statistický (stochastický) – strojové učení na referenčním treebanku
  - pravidlový – formální gramatika, popis frází a pravidla jejich spojování
- datové struktury – **závislostní/složkové stromy**
- automatická, poloautomatická, ruční anotace

K čemu to potřebujeme?

- další rovina popisu jazyka v NLP
- **treebank** – referenční data pro automatické nástroje
- synchronní (i diachronní) studie, vazba na slovesnou valenci a sémantickou rovinu
- frekvenční studie – SYN2015
- **navazující aplikace**, např. vývoj pravopisného a gramatického korektoru, aktuální členění věty (téma a réma), koreferenční vztahy (anafora a katafora), dialogové systémy
- **čeština** – jeden z nejobtížnějších jazyků – flexe a volný slovosled

Syntaktická analýza – Praha

- historické pozadí
  - lingvistický strukturalismus, Pražská škola
  - Pražský lingvistický kroužek (1926, Mathesius, Jakobson, Trnka)
- funkčně generativní popis (Functional Generative Description, FGP, Sgall, 60. léta)
  - závislostní syntax
  - hloubková (tektogramatická) struktura
  - formální popis aktuálního členění věty

Syntaktická analýza – Praha

- ÚFAL MFF UK
  - *https://ufal.mff.cuni.cz/pdt3.5*
- PDT 1.0–3.5 (*Prague Dependency Treebank, Pražský zavislostní korpus*)
- ruční anotace
- rovina anotace:
  - slovní
  - morfologická (2 mil. slovních jednotek)
  - syntaktická (analytická; 1,5 mil. slovních jednotek)
  - sémantická (tektogramatická; 0,8 mil. slovních jednotek)
  - aktuální členění věty, koreferenční vztahy, MWEs, analýza diskurzu
- teoreticky závislý, určen pro strojové učení

Syntaktická analýza – Brno

- CZPJ FI MU, syntaktické analyzátory
- **SYNT** – A. Horák, formální popis gramatiky (metagramatika, pravidla), složkové stromy
- http://nlp.fi.muni.cz/projekty/wwwsynt/
- **SET** – V. Kovář, pravidlový systém založený na vzorech, identifikace částí věty, složkové a závislostní stromy, keře (bush), přepíná mezi pozičním a atributivním systémem
  - nominální fráze (NP)
  - verbální fráze (VP)
  - koordinace (COORD)
- https://nlp.fi.muni.cz/projekty/set/wwwset.cgi/first_page