Počítačové zpracování češtiny

Syntaktická analýza

Daniel Zeman

http://ufal.mff.cuni.cz/course/popj1/

Syntaktická rovina

- Vztahy mezi větnými členy.
- Větný člen je pro nás slovo (tj. též interpunkce).
  - Z praktických důvodů:
    - Snadno rozpoznatelné.
    - Jednotka předcházející (morfologické) úrovně zpracování.
    - Definováno ve většině jazyků stejně, na rozdíl od větného členu.
  - Na druhou stranu:
    - Musíme definovat i technické vztahy uvnitř skutečného větného členu (pomocná slovesa, předložky apod.)
  - Ale některé syntaktické korpusy slova spojují či dělí.

Umístění syntaktické roviny

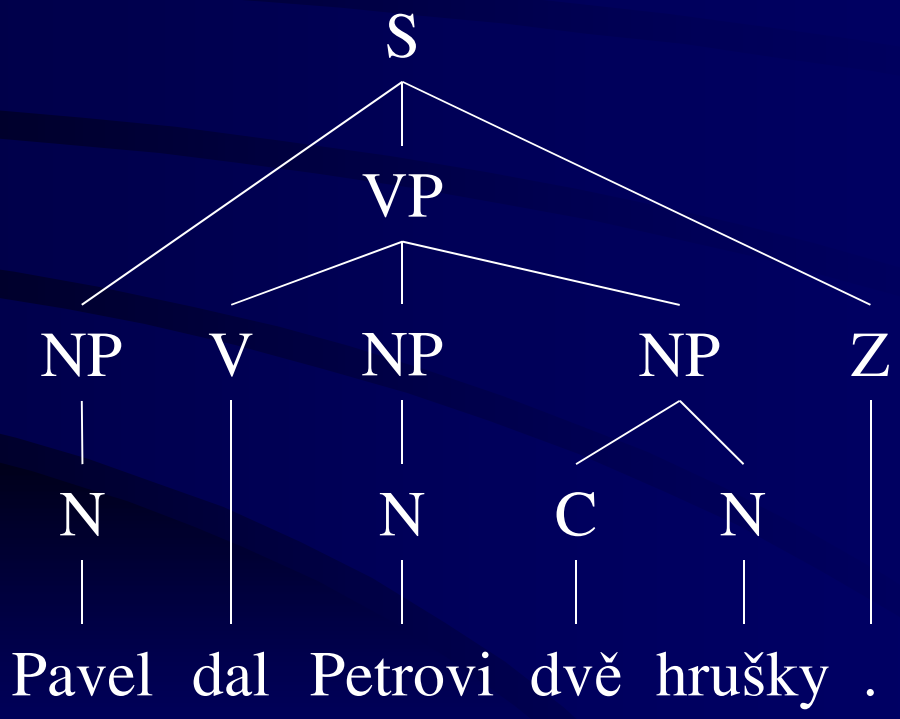
- Mezi morfologií a významem.
- Morfologie poskytuje / vyžaduje:
  - lemmata (je čas vytáhnout ze slovníku syntaktické informace)
  - značky (slovní druh a morfologické kategorie)
  - slovosled (nyní začíná být důležitý)
- Typicky nejednoznačný vstup
  - víceznačný výsledek morfologické analýzy.
- Typicky nejednoznačný výstup
  - více syntaktických struktur pro jednu větu (více interpretací věty).

Syntaktická struktura

- V různých teoriích má různý tvar
- Typicky nějaký strom
  - Frázový (složkový) strom
  - Závislostní strom

Příklad frázového stromu

- ((Pavel (dal Petrovi (dvě hrušky))) .)



Příklad závislostního stromu

- [#,0] ([dal,2] ([Pavel,1], [Petrovi,3], [hrušky,5] ([dvě,4])), [.,6])

Slova a fráze

- Slovo
  - nejmenší jednotka na syntaktické rovině
  - pomocná (funkční) slova (např. *a* v koordinaci *Pavel a Petr*, *být* ve složených slovesných tvarech *zkoušel jsem to*, *budu to zkoušet*)
  - významová (autosémantická) slova (např. *pes*; *být* ve větě *myslím, tedy jsem*)
- Fráze
  - skládá se ze slov a/nebo jiných frází (z **bezprostředních složek**)

Slova

- Idiomy
  - Pevné, neoddělitelné fráze se mohou chovat jako jedno slovo (např. složené předložky jako *na rozdíl od někoho*, cizí pojmenované entity jako *Rio de Janeiro*, ustálená spojení jako *být z něčeho na větvi*).
    - Zvláštní skloňování
    - Význam celku nelze složit z významů částí
    - Někdy tokenizace: Rio_de_Janeiro
- Vztahy k ostatním slovům
  - Slovník je zásobárna informací o slovech a vztazích mezi nimi.
    - Subkategorizace sloves (nevyžadují předmět, vyžadují a jaký…).
    - Významové rozlišení (podstatné jméno má barvu, má velikost, může být podmětem těch a těch sloves…).

Zaměnitelnost frází

- Frázi můžeme nahradit jinou frází stejného druhu.
  - Můžeme ji nahradit její hlavou
  - Vychází z představy, že věta je generována po frázích

⇒ Fráze *x, y, z* mohou být bezprostředními složkami větší fráze *f*, pouze pokud k sobě mají vztah. Konkrétní případy vymezuje konkrétní frázová gramatika.

  - Příklad: věta *„To je ten muž, o kterém jsem mluvil.“* Část *„muž, o kterém“* není jmenná fráze, protože ji nemůžeme nahradit jinou jmennou frází, např. *muž: „*To je ten muž jsem mluvil.“*

Fráze

- Fráze
  - Posloupnost bezprostředních složek (slov nebo frází).
  - Někdy nemusí být souvislá. Např. věta „*Soubor se nepodařilo otevřít.*“ obsahuje frázi „*otevřít soubor*“.
- Druhy frází podle druhu hlavního slova — hlavy
  - Jmenná (substantivní) fráze: *nová kniha mého dědečka*
  - Adjektivní fráze: *zbrusu nový*
  - Adverbiální (příslovečná) fráze: *velmi špatně*
  - Předložková fráze: *ve třídě*
  - Slovesná fráze: *chytit míč*

Jmenné fráze

- Hlavou je podstatné jméno nebo substantivní zájmeno.
  - *voda*
  - *ta kniha*
  - *nové nápady*
  - *dva milióny obyvatel*
  - *jedna malá vesnice*
  - *největší pohyb cen od druhé světové války během jednoho roku*
  - *operační systém, který navzdory veškerému úsilí našeho správce padá příliš často*
  - *on*
  - *kdokoli*

Adjektivní fráze

- Hlavou je přídavné jméno.
- Jednoduché ADJP jsou velmi časté, složené jsou řídké.
  - *starý*
  - *velmi starý*
  - *opravdu velmi starý*
  - *pětkrát starší než nejstarší slon v naší ZOO*
  - *jist, že tam bude první*

Zájmena

- Podobné chování jako podstatná jména (substantivní).
  - Osobní zájmena (*já*, *ty*, *oni*, *se*).
  - Některá tázací, vztažná, neurčitá a záporná (*kdo*, *co*, *někdo*, *něco*, *nikdo*, *nic*).
- Podobné chování jako přídavná jména (adjektivní).
  - Přivlastňovací zájmena (*můj*, *tvoji*, *jejich*, *svá*).
  - Ukazovací zájmena (*ten*, *ta*, *tamti*, *tyto*).
  - Některá tázací, vztažná, neurčitá a záporná (*který*, *jaký*, *čí*, *některý*, *lecjaký*, *žádná*).

Číselné fráze

- V češtině není vždy jasné, zda má být hlavou číslovka, nebo počítaná jmenná fráze.
  - Číslovka dědí rod počítaného jména. Jméno dostává číslo (jednotné / množné) podle číslovky.
    - *jeden muž, jedna žena, jedno dítě*
    - *dva muži, dvě ženy, dvě děti*
  - Číslovka určuje pád počítaného jména.
    - *pět mužů*
  - Číslovka i počítané jméno mají pád vyžadovaný předložkou nebo slovesem.
    - *pěti ženami*

Číselné fráze

- Podobné chování jako přídavná jména.
  - Základní číslovky 1 až 4 (*tři banány*).
  - Řadové číslovky (*čtyřicátý čtvrtý závodník*).
  - Některé druhové číslovky (*čtvery hodiny*, *jedni lidé*).
- Podobné chování jako podstatná jména.
  - Základní číslovky 5 a více v 1., 4. a 5. pádě.
  - Některé druhové číslovky (srov. *sedmero krkavců / sedm krkavců / hejno krkavců / přílet krkavců*).
- Podobné chování jako příslovce.
  - Násobné číslovky (*pětkrát*).
  - Řadové stažené s předložkou (*poprvé*).

Adverbiální fráze

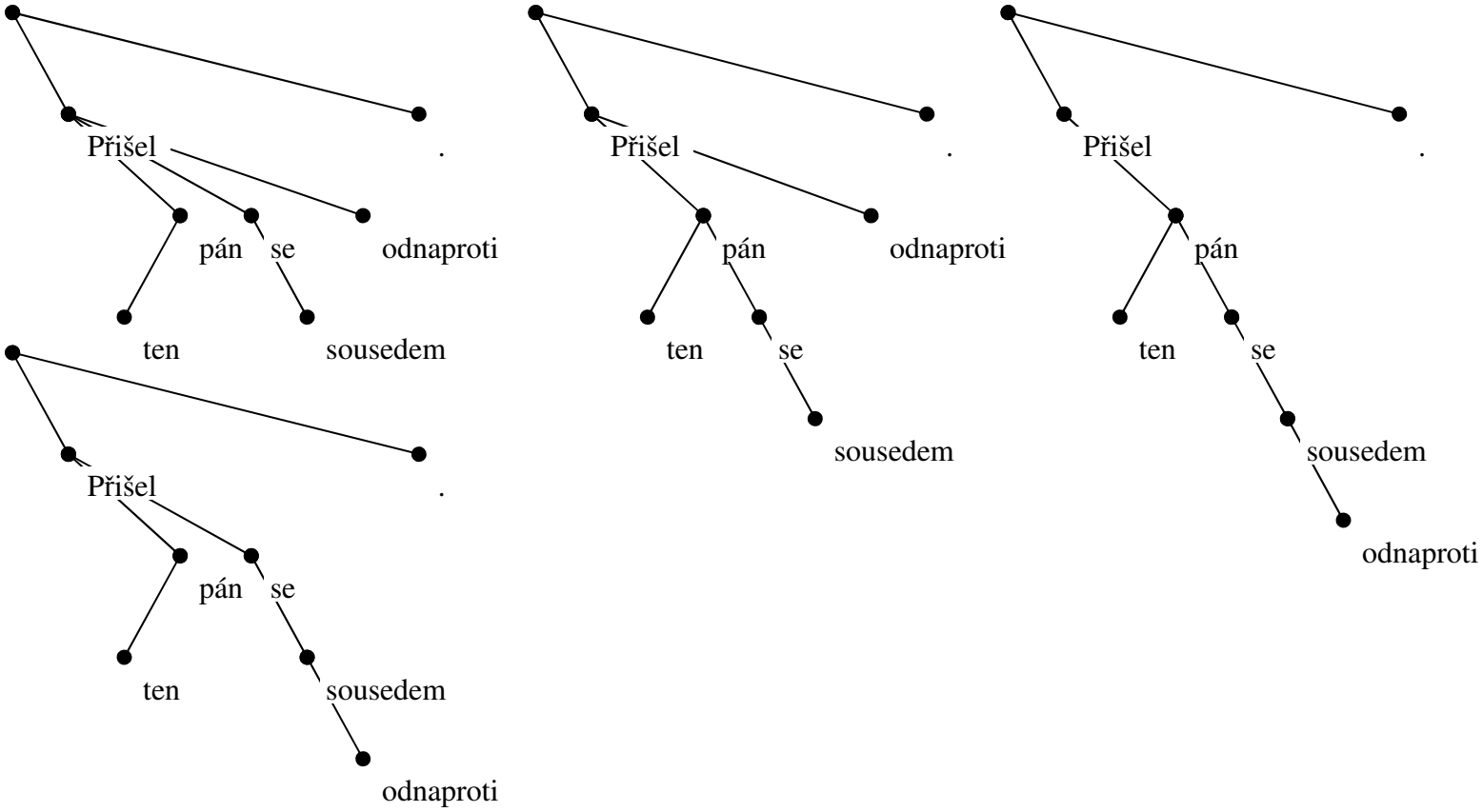
- Hlavou je příslovce.
  - *rychle*
  - *mnohem více*
  - *jak*
  - *hlasitěji, než si dovedete představit*
  - *včera*

Předložkové fráze

- Hlavou je předložka (protože určuje pád, ve kterém musí být zbytek fráze).
- Často podobná funkce jako adverbiální fráze (příslovečné určení) nebo jmenná fráze (předmět slovesa).
  - *v centru města*
  - *v náhodu*
  - *kolem páté*
  - *k lepším zítřkům*
  - *až do situace, kde nikdo z nich nemohl couvnout*
  - *vzhledem k jeho nezletilosti*

Předložkové fráze

- *„Přišel ten pán se sousedem odnaproti.“*



Předložkové fráze

- Anglický příklad:
  - *I saw the man with a telescope.*
    1. *Viděl jsem ho dalekohledem.*
    2. *Viděl jsem ho s dalekohledem.*

Předložkové skupiny a syntaktické nejednoznačnosti

- *V **letech** 1991 – 1993 jsem **absolvovala kurzy** řízení a marketingu na Collège Bart v kanadském Québecu.*

  - *absolvovala na Collège Bart*
  - *kurzy na Collège Bart*
  - *řízení a marketingu na Collège Bart*
  - *marketingu na Collège Bart*
  - *Collège Bart v Québecu*
  - *marketingu v Québecu...*

Předložkové skupiny a syntaktické nejednoznačnosti

- *V letech 1991 – 1993 jsem absolvovala kurzy řízení a marketingu na Collège Bart v kanadském Québecu.*

  - *absolvoval (kurzy (řízení a market)) (na Bartu)*
  - *absolvoval (kurzy (řízení a market) (na Bartu))*
  - *absolvoval (kurzy ((řízení a market) (na Bartu)))*
  - *absolvoval (kurzy (řízení a (market (na Bartu))))*
  - *… ((na Bartu) (v Québecu))*
    - Je Bart v Québecu, nebo Québec na Bartu?

Fráze s funkcí předložky

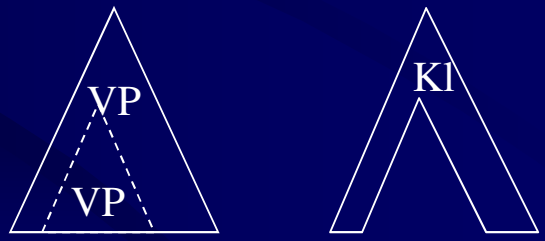
- Jsou hlavou předložkové fráze. Často se pojí s genitivem.
  - *prostřednictvím GEN*
  - *počátkem GEN*
  - *díky DAT*
  - *na základě GEN*
  - *na rozdíl od GEN*
  - *v rámci GEN*
  - *v průběhu GEN*
  - *v souladu s INS*
  - *do souladu s INS*
  - *s pomocí GEN*

Slovesné fráze

- Hlavou je podtržené sloveso.
  - *prší*
  - *mohl by vůbec spatřit pana prezidenta*
  - *, proč jsme tolik zmokli*
  - *Prosím!*
  - *v neděli byl převezen do nemocnice*
  - *začalo sněžit*
  - *zakazuje kouřit v této místnosti*
  - *dej Pavle ty korále, co jsme přivezli z dovolené v Maroku*
  - *soubor se nepodařilo otevřít*

Klauze

- Část věty, která obsahuje právě 1 přísudek, např.:
  - *Běží liška k táboru.*
  - *, že máte pravdu.*
  - přísudek nemusí být vyjádřen na povrchu, např. nadpisy: *Senzační odhalení pozadí (skandálu).*
- Není totéž, co slovesná fráze (VP).
  - Součástí VP mohou být menší VP.
  - Do klauze mohou být vnořené jiné klauze, které ale nejsou její součástí.



Klauze a věty

- Klauze
  - jednoduchá věta nebo věta v souvětí
  - např. *Běží liška k Táboru.* nebo *„že máte pravdu“*.
- Věta
  - jednoduchá věta i souvětí
  - může se skládat z jedné až několika klauzí
  - např. *Běží liška k Táboru.* nebo *Zjistil jsem, že máte pravdu*.

Klauze

- Predikativní (přísudková) funkce.
  - Jistá aktivita jistých podmětů (subjektů) a předmětů (objektů) v jistém čase za jistých podmínek.
- Hlavní klauze (hlavní věta).
  - Není závislá na jiné větší klauzi.
- Vnořená klauze (vedlejší věta).
  - Je závislá na jiné klauzi, v níž má jistou funkci (jako fráze).
- Funkce klauzí:
  - Tytéž co u frází, navíc některé speciální, např. přímá řeč.

Věty

- Skládají se z jedné nebo několika hlavních klauzí.
- Při několika hlavních klauzích obvykle koordinace klauzí (frází).
- V psaném textu začínají velkým písmenem (to se však může vyskytnout i uvnitř věty). Někdy začínají závorkou nebo uvozovkami.
- Končí tečkou, vykřičníkem nebo otazníkem (ale i tečka se může vyskytnout uvnitř věty). Někdy končí i závorkou nebo uvozovkami.
- Zda a kdy i středník a dvojtečka mohou ukončovat větu, závisí na pohledu, který zvolíme. Obvykle však je možné dívat se na ně jako na koordinační spojku.

Koordinace frází

- Hlavu zastupuje spojka, čárka apod.
- Koordinované fráze jsou obvykle stejného druhu.
  - *kuřata, slepice, králíci, kočky a psi*
  - *nová nebo ještě novější*
  - *rychle i kvalitně*
  - *došel k závěru, že nemá smysl nadále se skrývat, takže bychom ho tu dneska mohli slyšet*
  - *ve městě a na vsi*
  - *do a z Prahy*
  - *buď hned, nebo později*
  - *nejen v pondělí a ve středu, ale i zítra nebo pozítří*

Apozice frází

- Podobná jako koordinace, ale s jiným významem.
  - *Karel IV., císař římský a král český*
- Koordinace: několik různých členů plní danou funkci společně.
- Významově jen jeden člen, ale na povrchu má několik jmen či popisů.
  - *a nejvíce — 40 procent — je rodinných domků*
  - *faktorů, zejména však amortizace*
  - *správce — fyzická nebo právnická osoba, kterou určí vlastník domu*
  - *náklady a zvyšování daní — to jsou otázky, které…*

Apozice frází

- *veškeré jakostní znaky — jemnost, vlhkost, podmínky skladování a podobně*
- *doklad o zaplacení (útržek složenky nebo avízo při bezhotovostním převodu*
- *přesvědčen o jednom : je třeba mít vysoké cíle a nespokojit se s malými*
- *odbor dopravy městského úřadu , pan Jan Motyčka*

Elipsa

- **Elipsa** (výpustka), **elidovat** (vypustit).
- Fráze, která ve větě („na povrchu“) chybí, ačkoli je součástí významu („hloubkové struktury“).
- Často v rozhovorech: elidovaný člen je znám z kontextu.
  - *Koho jsi tam viděl? — Petra.* (Chybí sloveso.)
- V psaném textu často v koordinacích.
  - *Čeští a němečtí studenti se zúčastnili…* (Pravděpodobně nikdo nebyl Čech a Němec zároveň. Spíše to byli *čeští studenti a němečtí studenti*.)
  - *Slavia vede 4:0, zatímco Sparta jenom 3:2.* (Sloveso ve 2. části.)
- V češtině i jiných jazycích někdy systémová, např. vypouštění osobního zájmena, které je podmětem věty.
  - *Sedím.* (*já*)

Díry a nesouvislé fráze

- Složka (fráze) se přesunula z pozice, na které je očekávána.
- Běžně se o dírách (*gaps*) hovoří v souvislosti s angličtinou. U nás nepředstavují nic zvláštního, protože nemáme pevný slovosled. U nás se výraz *díra* používá odlišně (u neprojektivních konstrukcí)!
- V anglických otázkách a vztažných větách.
  - *Who do you work for <gap>$_{whom}$?*
  - *I don't know why we have got so much rain <gap>$_{why}$.*
  - *On Sundays, I usually work <gap>$_{on\ sundays}$ but I stay at home on Tuesdays.*
  - *the story he never wrote <gap>$_{the\ story}$*

Shrnutí frázového modelu

- Věta se dělí na fráze (složky).
- Fráze se může dělit na ještě menší fráze.
- Největší fráze je celá věta.
- Nejmenší fráze jsou slova.
- Fráze mají jména podle toho, jakého jsou druhu.

Shrnutí frázového modelu

- Frázi je možné nahradit jinou frází téhož druhu. Speciálně, lze ji nahradit jednou její bezprostřední složkou (hlavou).
  - Souvislost s generováním věty.

⇒ Fráze x, y, z mohou být bezprostředními složkami větší fráze f, jen pokud spolu nějak souvisejí. To je však věcí návrhu konkrétní frázové gramatiky.

  - Příklad: věta *„To je muž, o kterém jsem mluvil.“* Část *„muž, o kterém“* není celá jmenná fráze, protože ji není možné nahradit jinou jmennou frází, např. *muž*: *„*To je muž jsem mluvil.“*

Souvislost frázového modelu s bezkontextovou gramatikou

- Frázová struktura odpovídá derivačnímu stromu v gramatice, která danou větu generuje / přijímá.
- Příklad:
  - S → NP VP (věta má podmět a přísudek)
  - NP → N (jmenná fráze je podstatné jméno)
  - VP → V NP (slovesná fráze je sloveso a předmět)
- Slovníková část gramatiky:
  - N → pán | hrad | muž | stroj | Petr | Pavel | … | pána | muže …
  - V → vidí | nese | bere | maže | kryje | kupuje | … | viděl | nesl …

Slovník

- Slovníková část ve skutečnosti může být řešena mimo gramatiku.
- Například neterminály nejnižší úrovně (hned nad terminály) jsou morfologické značky.
  - Potom je nejnižší patro frázového stromu řešeno morfologickou analýzou a značkováním.
  - Gramatika pak pracuje jen s morfologickými značkami.

Rozšířený příklad gramatiky

- NP → N | AP N
- AP → A | AdvP A
- AdvP → Adv | AdvP Adv

- $NP_{nom} \rightarrow N_{nom}$
- $NP_{nom} \rightarrow AP_{nom}\ N_{nom}$
- $NP_{nom} \rightarrow N_{nom}\ NP_{gen}$

- $NP_{gen} \rightarrow N_{gen}$
- $NP_{gen} \rightarrow AP_{gen}\ N_{gen}$
- $NP_{gen} \rightarrow N_{gen}\ NP_{gen}$

- N → pán | hrad | muž | stroj …
- A → mladý | velký | zelený …
- Adv → velmi | včera | zeleně …

- N_{nom} → pán | hrad | muž …
- N_{gen} → pána | hradu | muže …
- N_{dat} → pánovi | hradu | muži …
- N_{acc} → pána | hrad | muže …
- N_{voc} → pane | hrade | muži …
- N_{loc} → pánovi | hradu | muži …
- N_{ins} → pánem | hradem …

Rozšířený příklad gramatiky

- $VP \rightarrow VP_{povinné}$
- $VP \rightarrow VP_{povinné}\ VP_{volitelné}$

- $VP_{povinné} \rightarrow V_{intr}$
- $VP_{povinné} \rightarrow V_{trans}\ NP_{acc}$
- $VP_{povinné} \rightarrow V_{bitr}\ NP_{dat}\ NP_{acc}$
- $VP_{povinné} \rightarrow V_{mod}\ VINF$

- $VP_{volitelné} \rightarrow AdvP_{místo} \mid AdvP_{čas}$
  …

- $V_{intr} \rightarrow$ šedivět | brzdit | krást …
- $V_{trans} \rightarrow$ koupit | ukrást …
- $V_{bitr} \rightarrow$ dát | půjčit | poslat …
- $V_{mod} \rightarrow$ moci | smět | muset …

- … (desítky až stovky rámců)

Unifikační gramatika

- Alternativa ke štěpení neterminálů
- Místo bezkontextových pravidel:
  - $NP_{nom} \rightarrow AP_{nom}\ N_{nom}$
  - $NP_{gen} \rightarrow AP_{gen}\ N_{gen}$
  - $NP_{dat} \rightarrow AP_{dat}\ N_{dat}$
  - $NP_{acc} \rightarrow AP_{acc}\ N_{acc}$
  - $NP_{voc} \rightarrow AP_{voc}\ N_{voc}$
  - $NP_{loc} \rightarrow AP_{loc}\ N_{loc}$
  - $NP_{nom} \rightarrow AP_{nom}\ N_{nom}$
- Unifikační pravidlo:
  - NP → AP N := [case = AP^case # N^case]

Syntaktická analýza (parsing)

- Automatické metody nalezení syntaktické struktury věty.
  - Symbolické metody: vyžadují frázovou gramatiku nebo jiný popis struktury jazyka. Pak: chart parser.
  - Statistické metody: vyžadují textový korpus se syntaktickými strukturami (tzv. treebank — stromová banka).
  - Kombinované metody: jednoduchá gramatika, nejednoznačnosti se řeší statisticky podle korpusu.
    - Chunking / shallow parsing („mělký rozbor“)

Syntaktická analýza podle bezkontextové gramatiky

- Hierarchie gramatik:
  - Noam Chomsky (1957): *Syntactic Structures*
- Několik klasických algoritmů.
  - CYK (Cocke-Younger-Kasami) … složitost $O(n^3)$
    - John Cocke („vynálezce“)
    - T. Kasami (1965), Bedford, MA, USA (jiný nezávislý „vynálezce“)
    - D. H. Younger (1967) (analýza složitosti)

    - Podmínka CYK: gramatika je v CNF (Chomského normální forma), tj. pravá strana jsou buď dva neterminály, nebo jeden terminál. (Lze snadno zařídit.)

Syntaktická analýza podle bezkontextové gramatiky

- **Chart parser**: CYK vyžaduje datovou strukturu pro udržování informace o rozpracovaných možnostech. Přelom 60. a 70. let: pro tento účel navržena struktura *chart* — přehled či diagram rozpracovaných a hotových složek věty.
- J. Earley (1968), disertace, Pittsburgh, PA, USA
  - Trochu jiná verze chart parsingu (analýzy s přehledem).

- Podrobněji o algoritmu chart parseru: viz dřívější přednášku o bezkontextových gramatikách a morfologii.

Frázový parsing v praxi

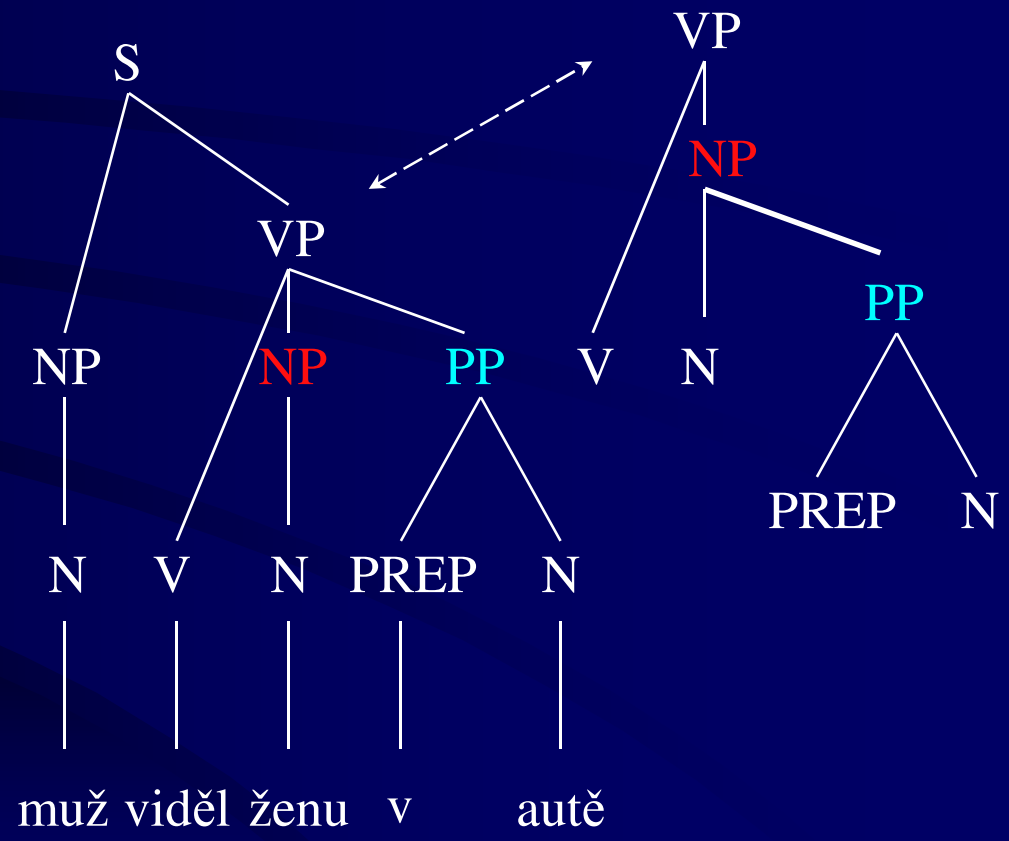
- Pravidlové parsery, např. Fidditch (Donald Hindle, 1983)
- Collinsův parser (Michael Collins, 1996–1999)
  - Pravděpodobnostní bezkontextové gramatiky, lexikalizované hlavy
  - Přesnost a úplnost na Penn Treebanku / Wall Street Journal data / Section 23 = 85%
  - Přepsáno do Javy Danem Bikelem ("Bikelův parser"), volně dostupný
- Charniakův parser (Eugene Charniak, NAACL 2000)
  - Inspirován statistickými modely maximální entropie
  - P ~ R ~ 89.5%
  - Mark Johnson: reranker => přes 90%
- Stanfordský parser (Chris Manning et al., 2002–2010)
  - Vyrábí složky i závislosti. P ~ R ~ 86.4%

Pravděpodobnostní bezkontextové gramatiky

- PCFG *(probabilistic context-free grammars)*
- Máme-li více možných analýz, chceme jim přiřadit váhy.
- Více možností se objeví tam, kde můžeme použít více pravidel se stejnou levou stranou.
- Myšlenka: pravděpodobnostní rozložení pravidel se stejnou levou stranou.
  - Příklad: gramatika povoluje **VP** → **V NP** i **VP** → **V NP PP**.
  - Vstupní věta rovněž umožňuje obě interpretace.
  - Ale víme (např.), že druhý způsob tvoření **VP** je častější:
    p(**V NP** | **VP**) = 0,3
    p(**V NP PP** | **VP**) = 0,7

Příklad nejednoznačné syntaktické analýzy

- S → NP VP
- VP → V NP PP
- VP → V NP
- NP → N
- NP → N PP
- PP → PREP N
- N → muž
- N → ženu
- N → autě
- V → viděl
- PREP → v



Pravděpodobnost derivačního stromu

- Obě fráze / analýzy jsou „gramatické“.
- Různé významy. Který je lepší v daném kontextu?
- Pravděpodobnostní bezkontextová gramatika:
  - Vztahy mezi rodičovskými uzly a dětmi.
  - Pravděpodobnost odvození, použití pravidla.
  - Pravděpodobnost celého derivačního stromu (r_i jsou pravidla gramatiky použitá ke generování věty S, jejíž analýzou je T):

$$p(T) = \prod_{i=1}^{n} p(r_i)$$

Předpoklady

- Použití pravidla je nezávislé na použití jiných pravidel ve větě (velmi silný a nepravdivý předpoklad).
- Nezávislost na kontextu okolních podstromů.
- Nezávislost na kontextu předků (vyšších úrovní).
- Nezávislost na umístění ve větě (slovosled) či ve stromu.

Pravděpodobnost pravidla

- Pravidlo r_i: A $\rightarrow \alpha$.
- Označme R_A množinu všech pravidel r_j, která mají na levé straně neterminál A.
- Na R_A definujme pravděpodobnostní rozložení:

$$\sum_{r \in R_A} p(r) = 1 \qquad 0 \leq p(r) \leq 1$$

- Jinými slovy:

$$p(r) = p(\alpha | A) \qquad r = A \rightarrow \alpha \qquad \alpha \in (N \cup T)^+$$

Odhad pravděpodobnosti pravidla

- Syntakticky označený korpus založený na bezkontextové gramatice (tedy ne např. závislostní korpus).

$$r = A \rightarrow \alpha_1 \alpha_2 \ldots \alpha_k \qquad p(r) = \frac{c(r)}{c(A)}$$

- Četnost použití pravidla: jak často se v korpusu objeví podstrom