Syntaktická analýza češtiny a její aplikace – ,,towards my thesis"

Vojtěch Kovář

Centrum zpracování přirozeného jazyka
Fakulta Informatiky, Masarykova Univerzita
Botanická 68a, 602 00 Brno

xkovar3@fi.muni.cz

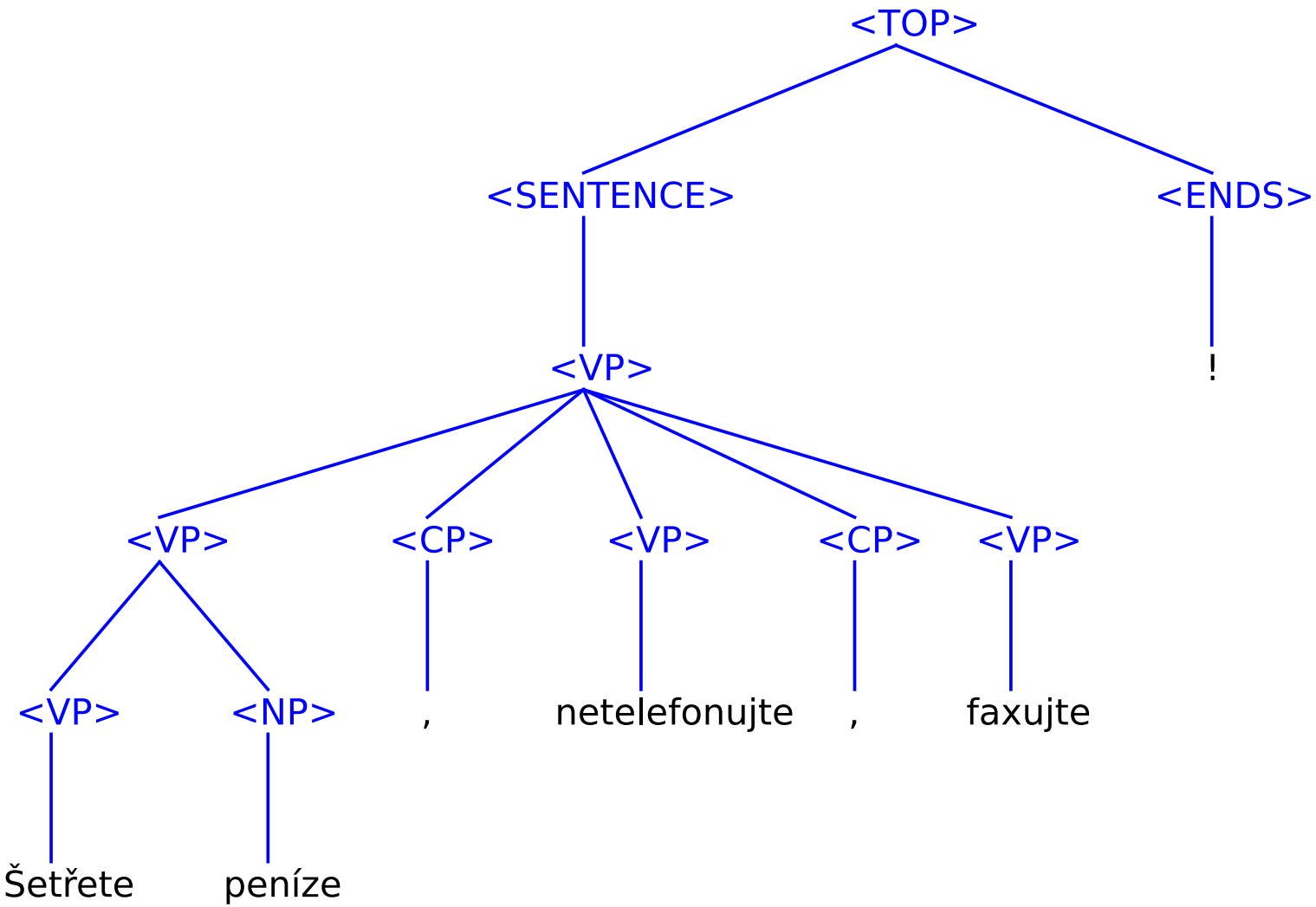
11.11.2010

Obsah

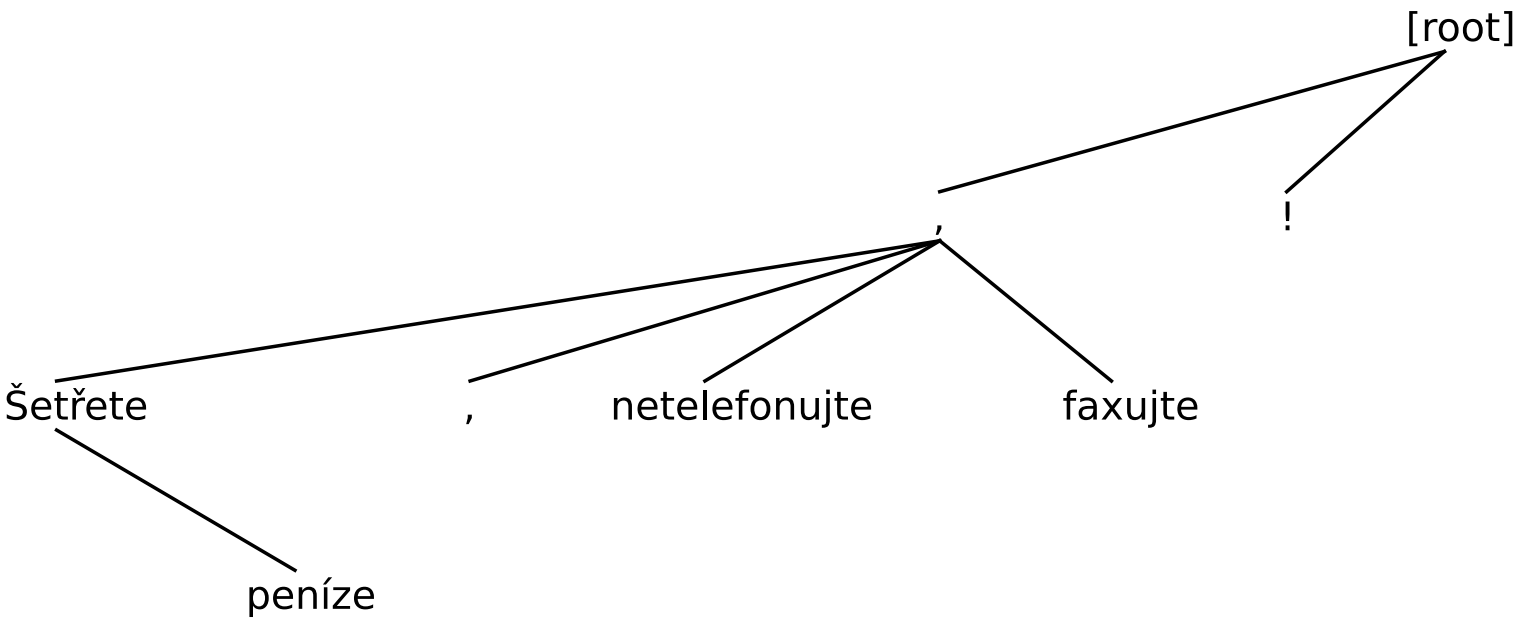
Syntaktická analýza přirozeného jazyka

- Co? Proč?
  - odhalení strukturních vztahů ve větě
  - hranice frází, závislosti...
  - ,,rozbor věty" na střední škole
  - základ pro pokročilejší analýzu věty
- Kódování syntaxe přirozeného jazyka
  - syntaktické stromy
  - složkové – odvození z CFG
  - závislostní – závislosti mezi slovy
  - hybridní – kombinace předchozích dvou
- Parciální syntaktická analýza
  - vyznačení hranic a typů frází v textu

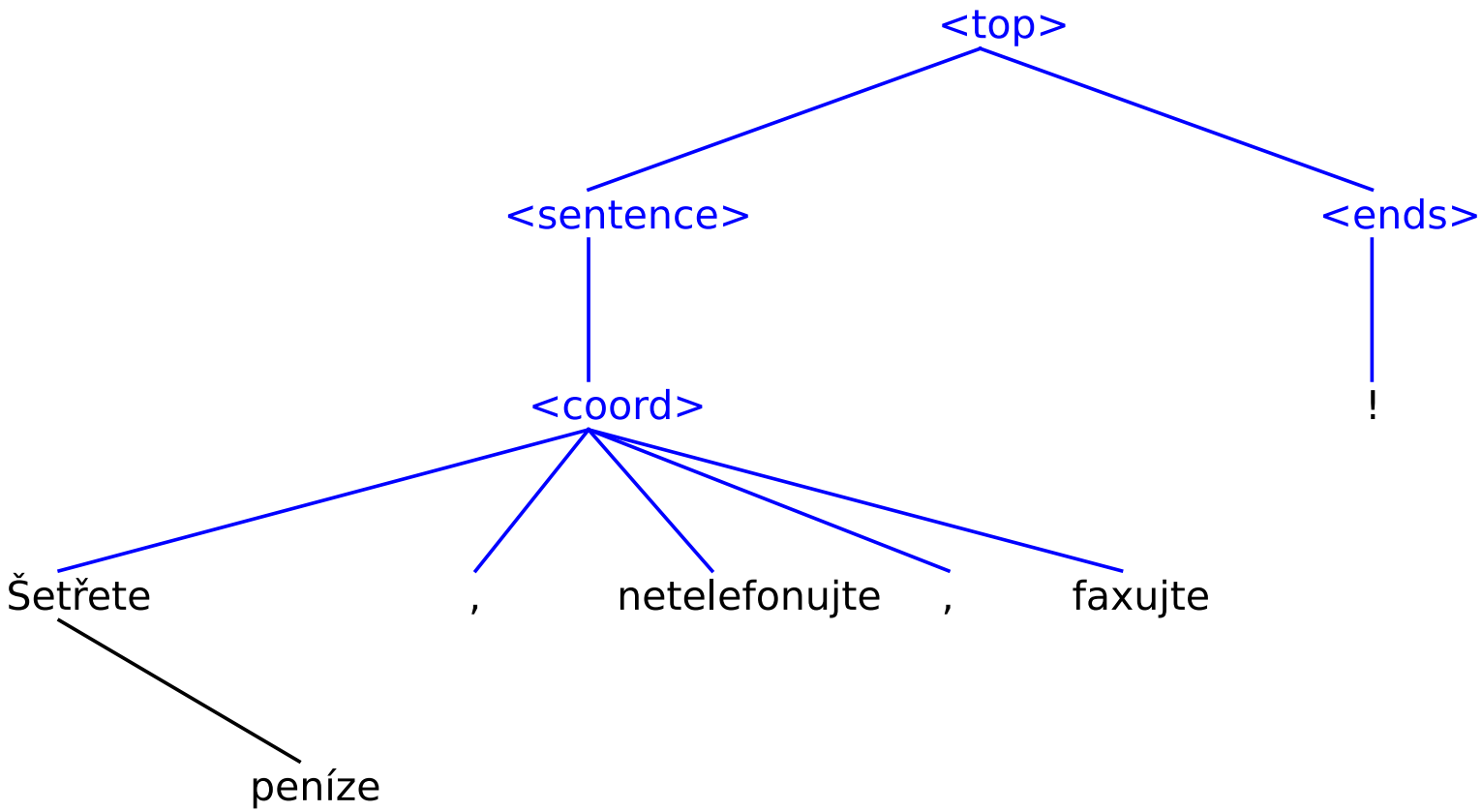
Věta: Šetřete peníze, netelefonujte, faxujte!



Věta: Šetřete peníze, netelefonujte, faxujte!



Věta: Šetřete peníze, netelefonujte, faxujte!



Automatická syntaktická analýza?

- Prerekvizity
  - rozdělení textu na slova (tokenizace)
  - morfologická analýza
  - → základní tvar, slovní druh, pád, číslo, rod...
- Principy automatických analyzátorů
  - pravidlové systémy
  - → ručně napsaná formální gramatika
  - → ručně napsaná sada pravidel
  - statistické systémy
  - → indukce jednoduchých pravidel z anotovaných dat
  - → hledání maximální kostry v grafu

Hodnocení kvality syntaktické analýzy

- Manuálně anotované korpusy
  - = velké soubory stromů vět
  - Penn Treebank (PTB)
  - Pražský závislostní korpus (PDT)
- Metriky podobnosti stromů
  - společné hrany
  - společné neterminály, cesty od kořene k listu
  - výstupem je ,,procento shody"
  - PARSEVAL, Leaf-ancestor assessment, precision, recall
- Hodnocení kvality analýzy
  - procento shody s daty v anotovaných korpusech

Nejlepší současné výsledky

- Angličtina
  - PARSEVAL (podobnost stromů, PTB)
  - → 92.1 % (McClosky, Charniak, and Johnson, 2006)
  - detekce jmenných frází (F-measure, PTB)
  - → 95.2 % (Shen and Sarkar, 2005)
- Čeština
  - závislostní přesnost (podobnost stromů, PDT)
  - → 86.3 % (Nakagawa, 2007)
  - detekce jmenných frází (precision, PDT – 1200 vět)
  - 93.1 % (Grác, Jakubíček and Kovář, 2010)

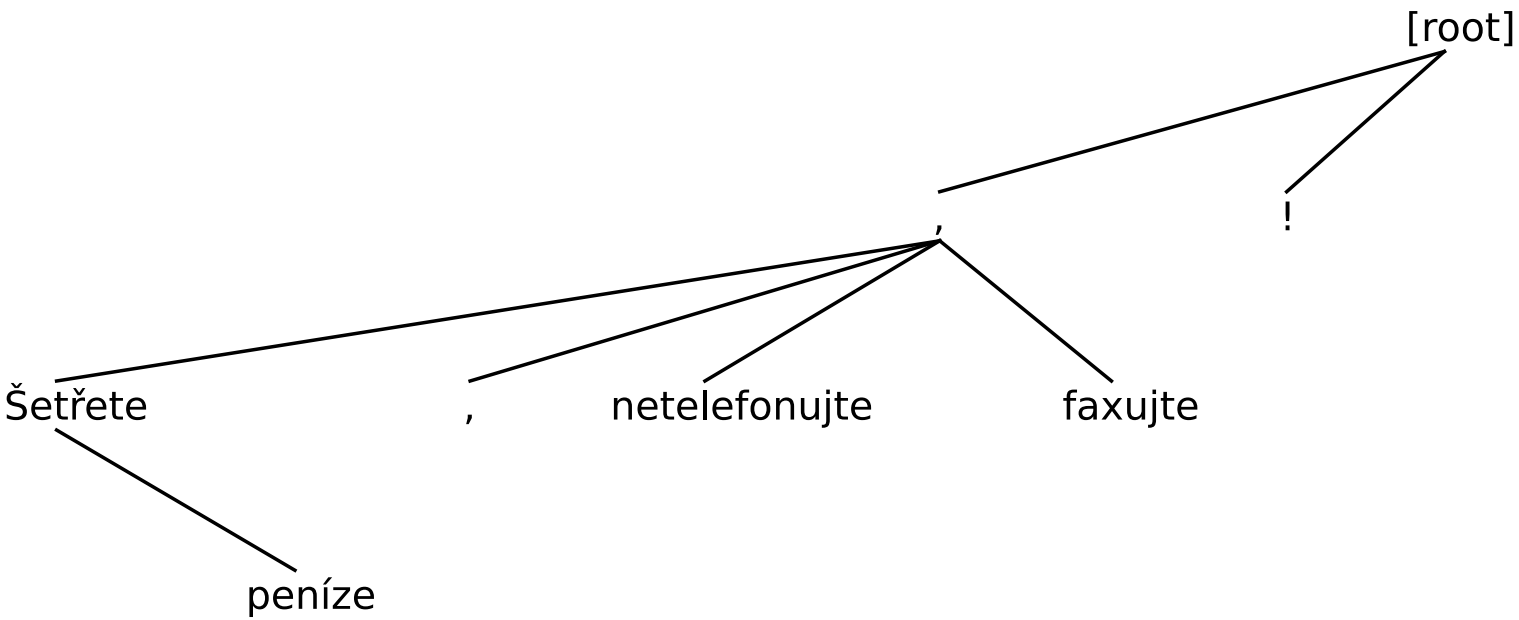
Problémy současné syntaktické analýzy

- Úspěšnost
  - uvedená čísla nejsou dostatečná
  - chyby analyzátorů jsou pro většinu aplikací limitující
- Metriky úspěšnosti
  - nejsou reprezentativní
  - nereflektují skutečnou využitelnost analyzátorů v reálných aplikacích
- Anotovaná data
  - složité struktury
  - chyby, nekonzistence, arbitrární konstrukce, výběr textů, ...
- Velmi aktuální diskuse

Problémy současné syntaktické analýzy (2)

- Zbytečné a neřešitelné problémy
  - ,,Karel mluvil o sexu s Britney Spears"
  - lidé nejsou bez dodatečných znalostí určit analýzu
  - ,,Letadlo spadlo do pole za lesem"
  - není důležité, kam se zavěsí zvýrazněná fráze

Věta: Šetřete peníze, netelefonujte, faxujte!



Cesta k řešení problémů – ,,towards the thesis"

- Zjednodušení konstrukce anotovaných dat
  - nyní: stovky stran manuálů pro anotátory
  - cíl: jednotky stran
- Vývoj reprezentativnějších měřítek kvality analýzy
  - nyní: metriky na stromech
  - cíl: ,,benchmarkové sady" založené na využití v aplikacích
  - $\longrightarrow$ detekce interpunkce, extrakce faktů, morfologická desambiguace
- Vývoj a zlepšení současných analyzátorů
  - nyní: optimalizace vzhledem k anotovaným datům
  - cíl: ,,application driven development"
  - principy YAGNI, KISS, worse is better, ...

Dosavadní výsledky, publikace

- Dosavadní výsledky
  - návrh a vývoj nového analyzátoru pro češtinu
  - identifikace častých chyb v anotovaných datech
  - návrh alternativní syntaktické anotace
- Publikace
  - Jakubíček, Horák, Kovář. Mining Phrases from Syntactic Analysis. (2009)
  - Kovář, Jakubíček. Prague Dependency Treebank Annotation Errors: A Preliminary Analysis. (2009)
  - Kovář, Horák, Jakubíček. Syntactic Analysis as Pattern Matching: The SET Parsing System. (2009)
  - Grác, Jakubíček, Kovář. Through Low-Cost Annotation to Reliable Parsing Evaluation (2010)