ijáček a slovosled

<NP n="S" g="M" c="5">Máchale</NP>

(REFLEXIVE c z) → (word=se|si|ses|sis c z)

(word=ti lemma=ty k=3 n=S c=3 p=2 x=P)

(INTER 3–4) → (COMMA 3–4) • (NP c=5) (COMMA)

Ondra Svoboda

O čem to bude

asi určitě o syntaxi ;-)

formální gramatika a slovosled

nový pravidlový parser

Gramatika (úvod)

„Tomáš učil Radku angličtinu.“

```
(S g n) →
   (NP c=1 g n              word=Tomáš)
   (k=5 m≠F g n             word=učil)
   (NP c=4                  word=Radku)
   (NP c=4                  word=angličtinu)
   (FULLSTOP                word=.)
```

Slovosledný model věty

Aleš Svoboda: České slovosledné pozice
z pohledu aktuálního členění (1984)

		upadlo		– slovesná skupina
Něco		upadlo		– větné členy
Něco	ti	upadlo		– příklonky
Něco	ti	upadlo,	Máchale	– vsuvky a spol.
Něco	ti	upadlo,	Máchale!	– interpunkce

Jak se rodí věty

budeme zítra dělat guláš

(lemma=být m=B) __ (k=5 m=F) __

↑ ↑

(CONSTITUENTS) (CONSTITUENTS)

S mamkou / ε – oznamovací věta

Proč – tázací věta

, že + oznamovací věta → obsahová vedlejší věta

, když – časová vedlejší věta

, (na) jehož NP – vztažná vedlejší věta

Oznamovací věty

uvaří

uvaří kávu

Uvaří ti kávu.

↑ ↑ ↑ ↑

Petr uvaří kávu

Petr ti uvaří kávu.

↑ ↑ ↑ ↑ ↑

, i když ji sám nepije,

Tázací věty

(Petr ti uvařil kávu?)

Uvařil ti (Petr) kávu?

Kdy ti (Petr) uvařil kávu?

Obsahové vedlejší věty

, že uvaří kávu.
, že ti uvaří kávu.
, že Petr uvaří kávu.
, že ti Petr uvaří kávu.

, že Petr ti uvaří kávu.

? , že uvaří ti kávu.

Gramatika (v praxi)

Něco ti upadlo, Máchale!

(k=5)

Něco ti upadlo, Máchale!

(CONSTITUENT)

Něco ti upadlo, Máchale!

(CLITICS)

Něco ti upadlo, Máchale!

(INTER)

Něco ti upadlo, Máchale!

(FULLSTOP)

Gramatika (v praxi)

Něco ti upadlo, Máchale!

(k=5)

(CONSTITUENT)

(CLITICS)

(INTER)

(FULLSTOP)

(CLAUSE) → (CONSTITUENT) (CLITICS) (k=5)
(INTER) (FULLSTOP)

Gramatika (v praxi)

(CLAUSE k=5 e=A g=N n=S a=P m=A) →

(INTER)

(CONSTITUENT) → (NP k=3 n=S c=1 word=Něco)

(INTER)

(CLITICS) → (INDIRECT_OBJ_PRONOUN)

→ (k=3 g n=S c=3 p=2 x=P word=ti)

(INTER)

(k=5 e=A g=N n=S a=P m=A word=upadlo)

(INTER)

(FULLSTOP)

Koordinace vedlejších vět?

... , že ti uvaří kávu, (že) přinese snídani do postele a (že) se celý den bude usmívat ...

(DEP_CLAUSE) → (DEP_CLAUSE) (CONJ) (ANOTHER_DEP_CLAUSE)

Yet another NIH parser

top-down chart parser podle Earleyho algoritmu, s unifikací atributů (pro gramatickou shodu)

expresivní a modulární systém pravidel

výstup pod kontrolou, přímo dostupný chart na postprocessing

NIH = Not Invented Here

Earleyho algoritmus

(INTER 3–4) → (COMMA 3–4) • (NP c=5 head) (COMMA)

predikce

(NP g n c=5 4–4) → • (k=1 g n c=5)

načtení terminálu

(NP g=M n=S c=5 4–5) →
(word=Máchale lemma=Máchal k=1 g=M n=S c=5 4–5) •

kombinace

(INTER 3–5) → (COMMA 3–4) (NP g=M n=S c=5 4–5 head) •
(COMMA)

Modulární pravidla pro češtinu

verbal_phrases → clauses, attributive_clauses, subordinate_clauses (zatím povrchově, bez rozlišení podle významu)

clitics (příklonky)

nominals, adverbials, constituents (jmenné a příslovečné fráze, další větné členy: podmětné/předmětné věty, fráze s infinitivem)

coordinations (všechny na jednom místě)

interpositions (vsuvky, vokativy, částice)

punctuation, optional_punctuation (povinné i nepovinné čárky a spol.)

vernacular (pravidla pro současný/hovorový jazyk)

ijáček (libreparser)

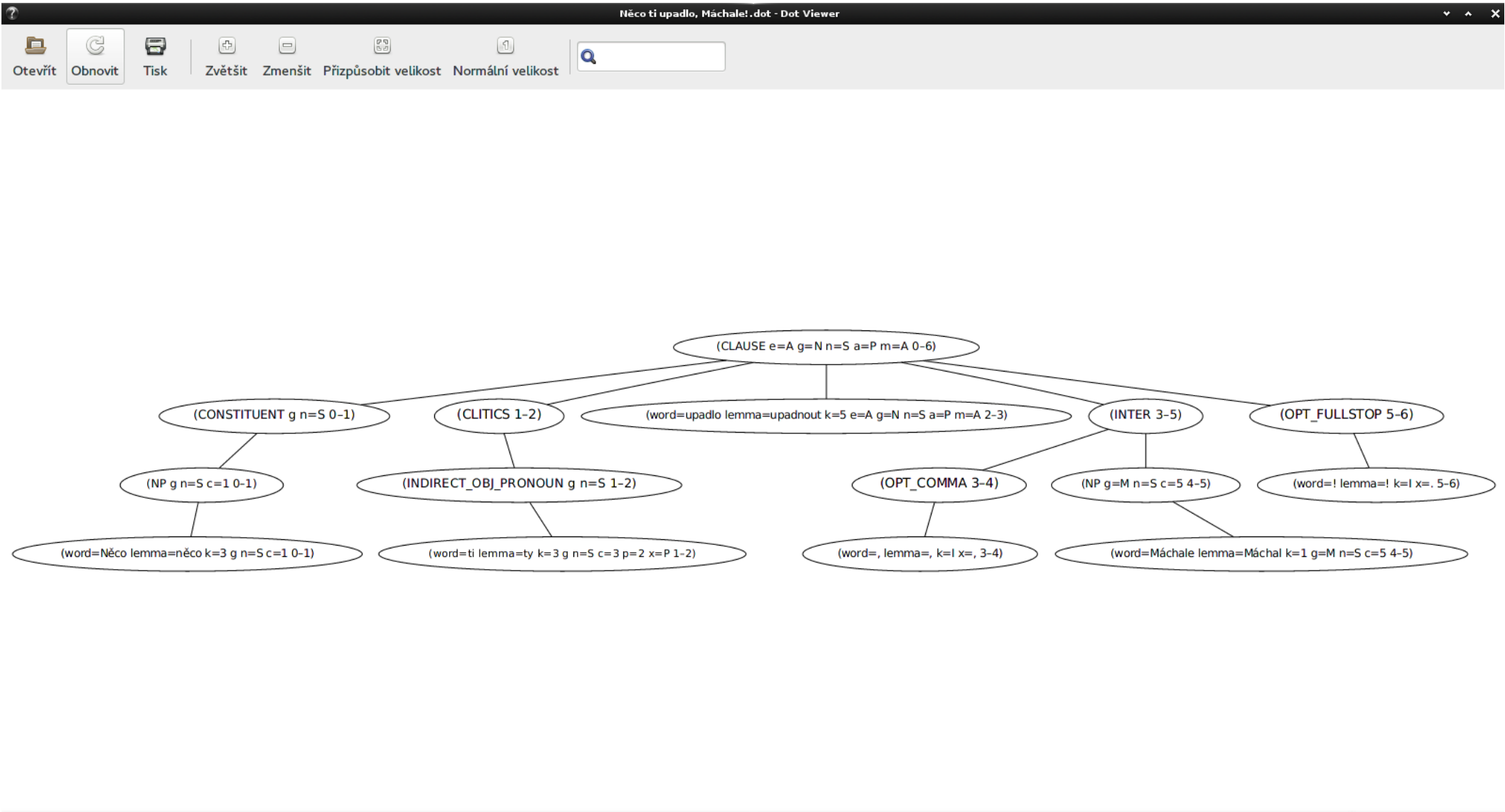
důraz na čitelnost kódu, podrobné informace o (probíhající) analýze → výukové účely

(plánovaná) možnost fallbacku na jiná pravidla po neúspěšné analýze

extrakce frází

rozšiřitelnost: další moduly gramatiky, jiné jazyky (i tagsety)

Grafický výstup (DOT)



Frázový výstup

```
<s seq="1" passed="yes">
<clause a="P" e="A" n="S" g="N" m="A">
<constituent n="S">
<np n="S" c="1">
Něco        něco            k3nSc1
</np>
</constituent>
<clitics>
<indirect_obj_pronoun n="S">
ti          ty              k3xPp2nSc3
</indirect_obj_pronoun>
</clitics>
upadlo      upadnout        k5eAaPmAgNnS
<g/>
<inter>
<opt_comma>
,           ,               kIx,
</opt_comma>
<np n="S" g="M" c="5">
Máchale Máchal  k1gMnSc5
<g/>
</np>
</inter>
<opt_fullstop>
!           !               kIx.
</opt_fullstop>
</clause>
</s>
```

Vyhodnocení a testování

zatím pokryto 27 % vět v DESAMu, proces běžel 8 hodin

self-check proti množině očekávaných hran → success rate

<s expected_edge="(CLAUSE e=A g n=S a=I m=I 0–5) → (word=Líbí lemma=líbit k=5 e=A g n=S p=3 a=I m=I 0–1) (CLITICS 1–4) (OPT_FULLSTOP 4–5) •">

v případě neúspěchu: vypisovat nejdelší neukončenou hranu?

Acknowledgements

Martin Kay (chart parsing)

Jay Earley (algortimus predikce, • načtení, kompletace 0–6)

James Allen (využití featur)

Vojta Kovář (jeho SET a IA161)

Aleš Svoboda (české slovosledné pozice)