Co nového ve zpracování MWE

Automatická identifikace

Společný workshop tří GAČRů

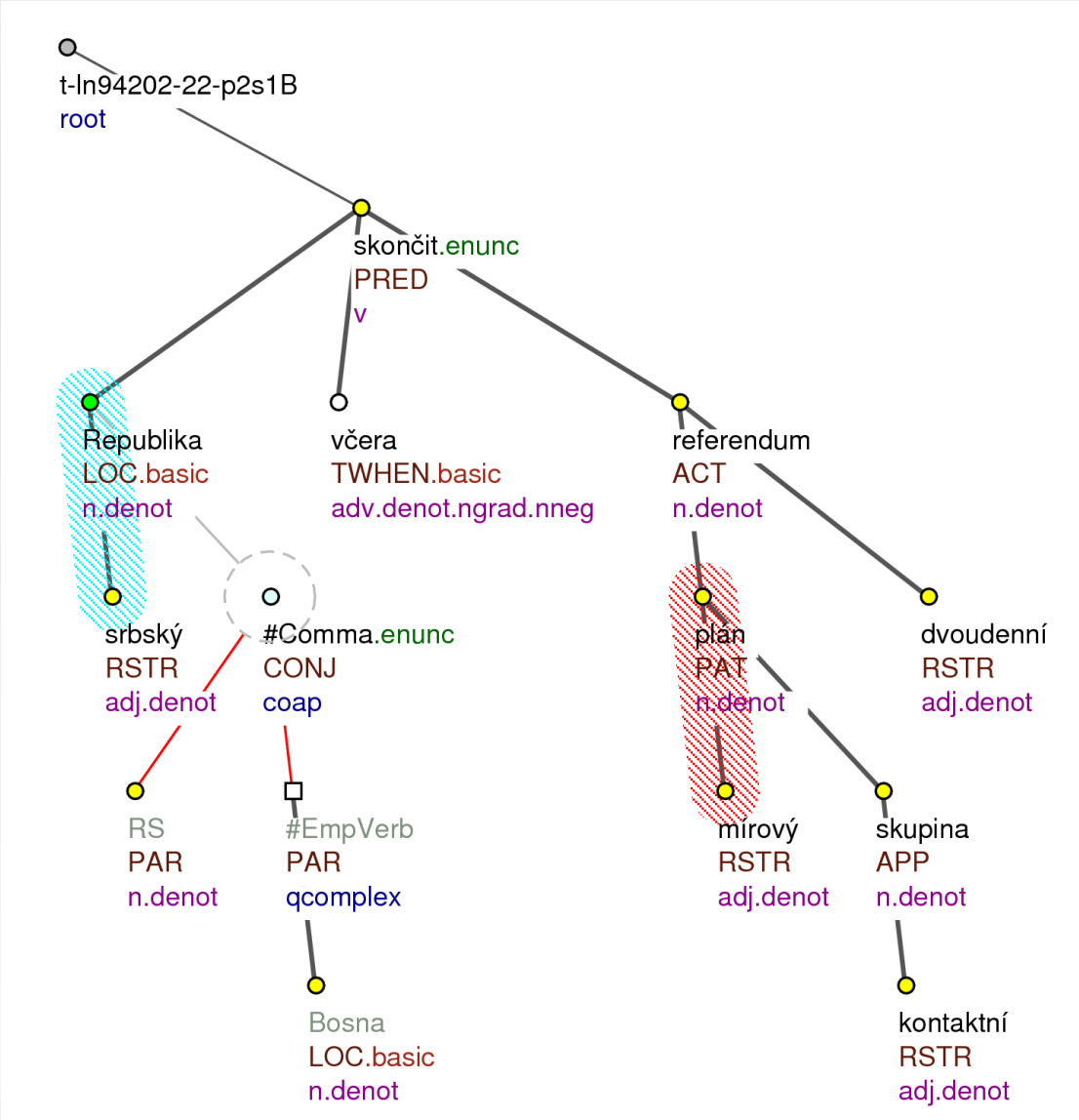
15. dubna 2013

Eduard Bejček

Víceslovné výrazy
Osnova

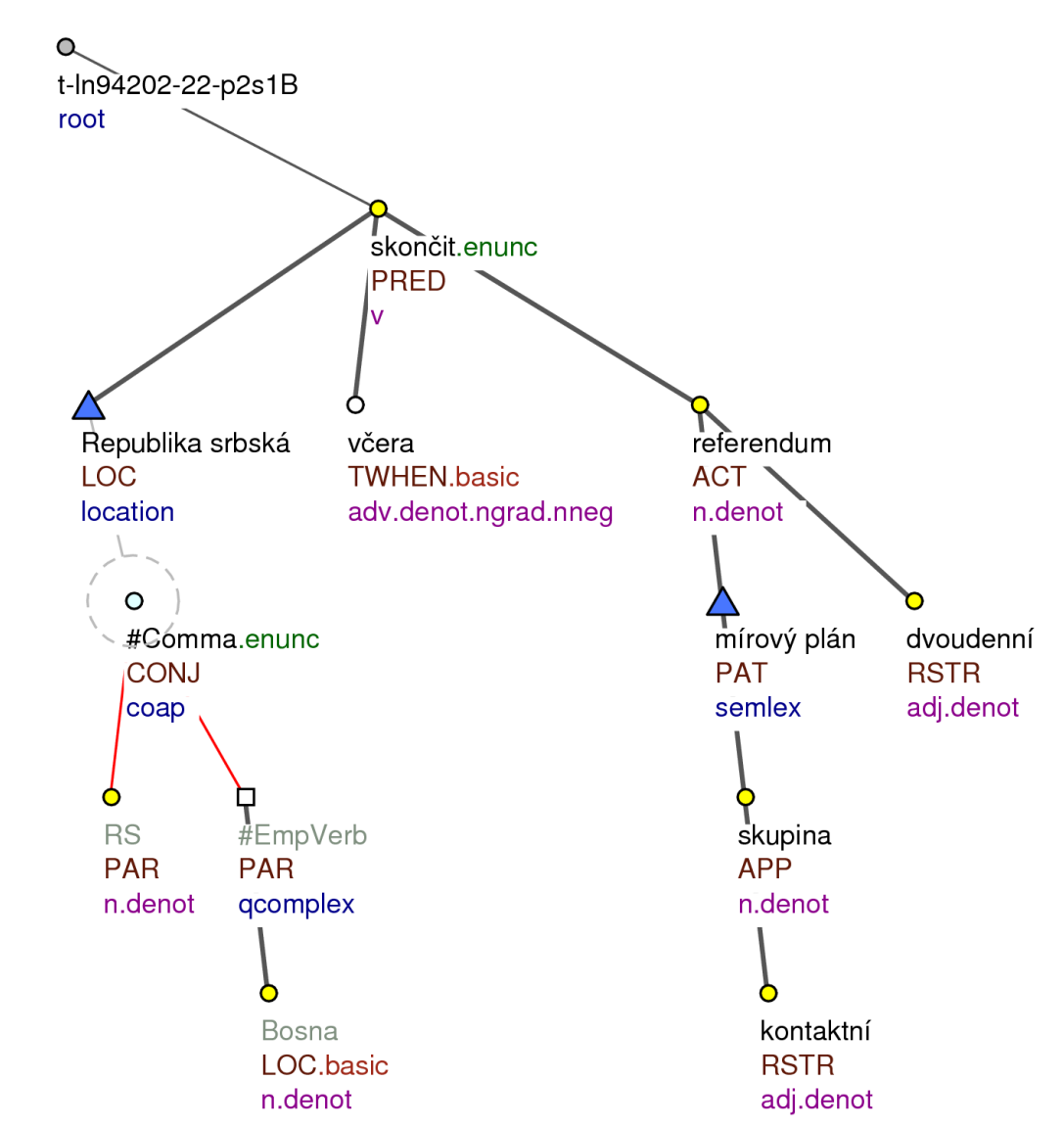
- Víceslovné výrazy (VV, MWE) v PDT 2.5
- Automatická identifikace
- Problémy
- (Úpravy slovníku)

PDT 2.5 – screenshot



V Republice srbské
(RS, v Bosně) včera skončilo
dvoudenní referendum
o mírovém plánu kontaktní
skupiny.

PDT 2.5 – screenshot



V Republice srbské
(RS, v Bosně) včera skončilo
dvoudenní referendum
o mírovém plánu kontaktní
skupiny.

Co jsou VV?

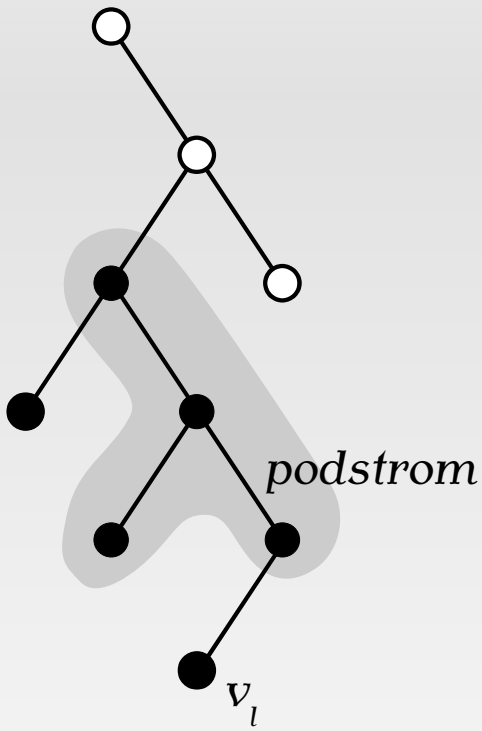
- řada kritérií – jen vodítka
- uloženy ve slovníku SemLex
- „když anotátor považoval za správné vložit do slovníku“
- **víc než jen kolokace**

- *detaily:*
  - komposicionalita — *neblahý konec* vs. *vysoká škola*
  - překlad — *high school*
  - substituovatelnost — *účetní poradce* vs. *účetní závěrka*
  - variovatelnost — **dopravní hřích*
  - odlučitelnost — **dopravní závažný přestupek*

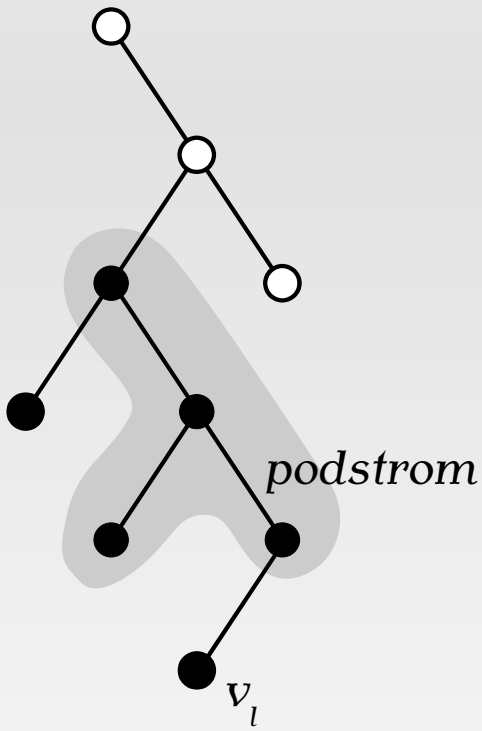
SemLex

- Slovník VV z celé t-roviny PDT
  - slovník má smysl – výběr netriviální
  - pouze víceslovné lexie – pojmenované entity nikoli
- uložena též stromová struktura („podstrom“)
  - jinak málo metadat
  - **předpoklad 1:** jeden VV = jedna struktura pro všechny výskyty
  - **předpoklad 2:** má-li něco strukturu VV, je to VV
- téměř 9 000 slovníkových položek

Stromová struktura v SemLexu



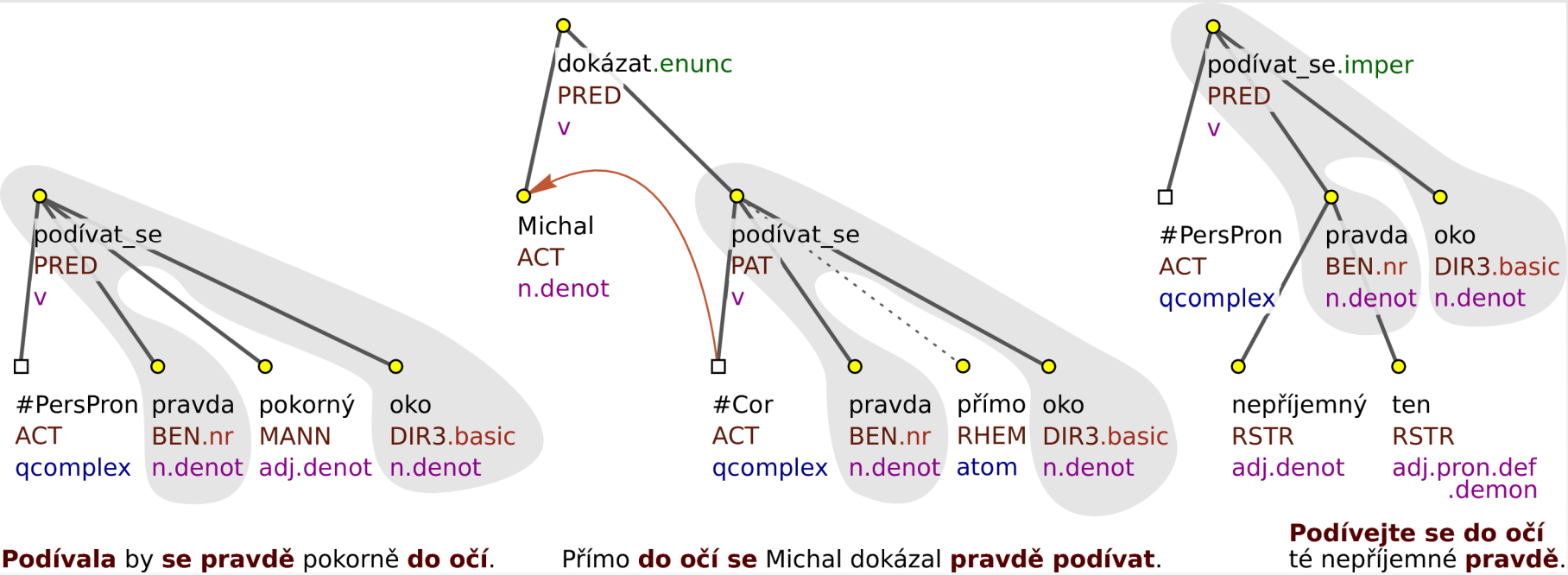
Stromová struktura v SemLexu



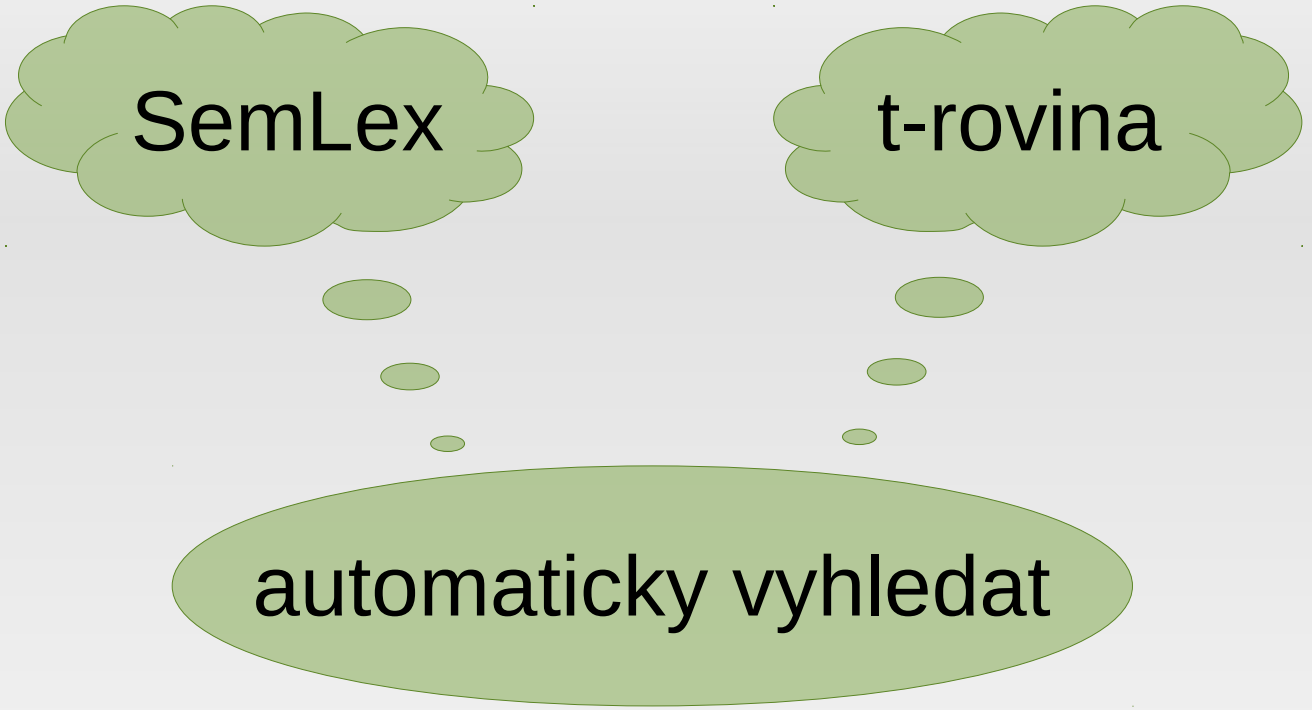
#root Nový ředitel peněžního odboru zmíněné banky podal resignaci.

Podstrom pro frazém

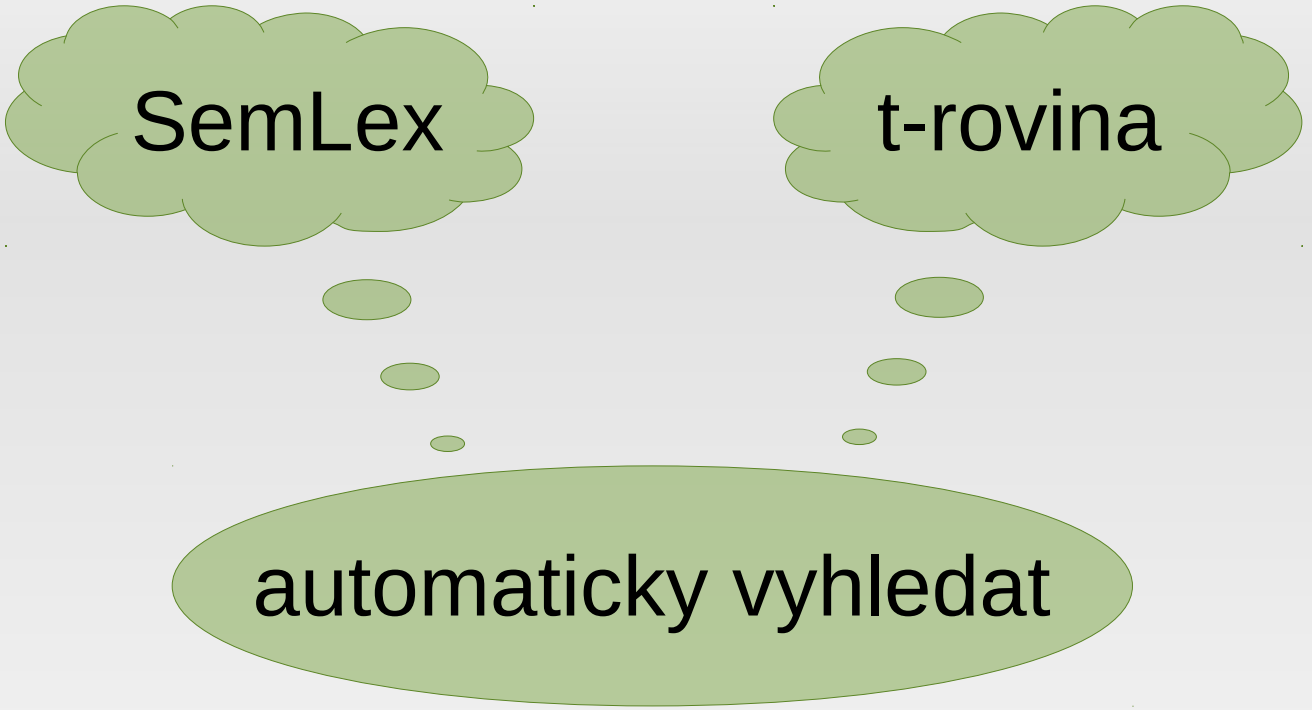
podívat se pravdě do očí



Automatická identifikace



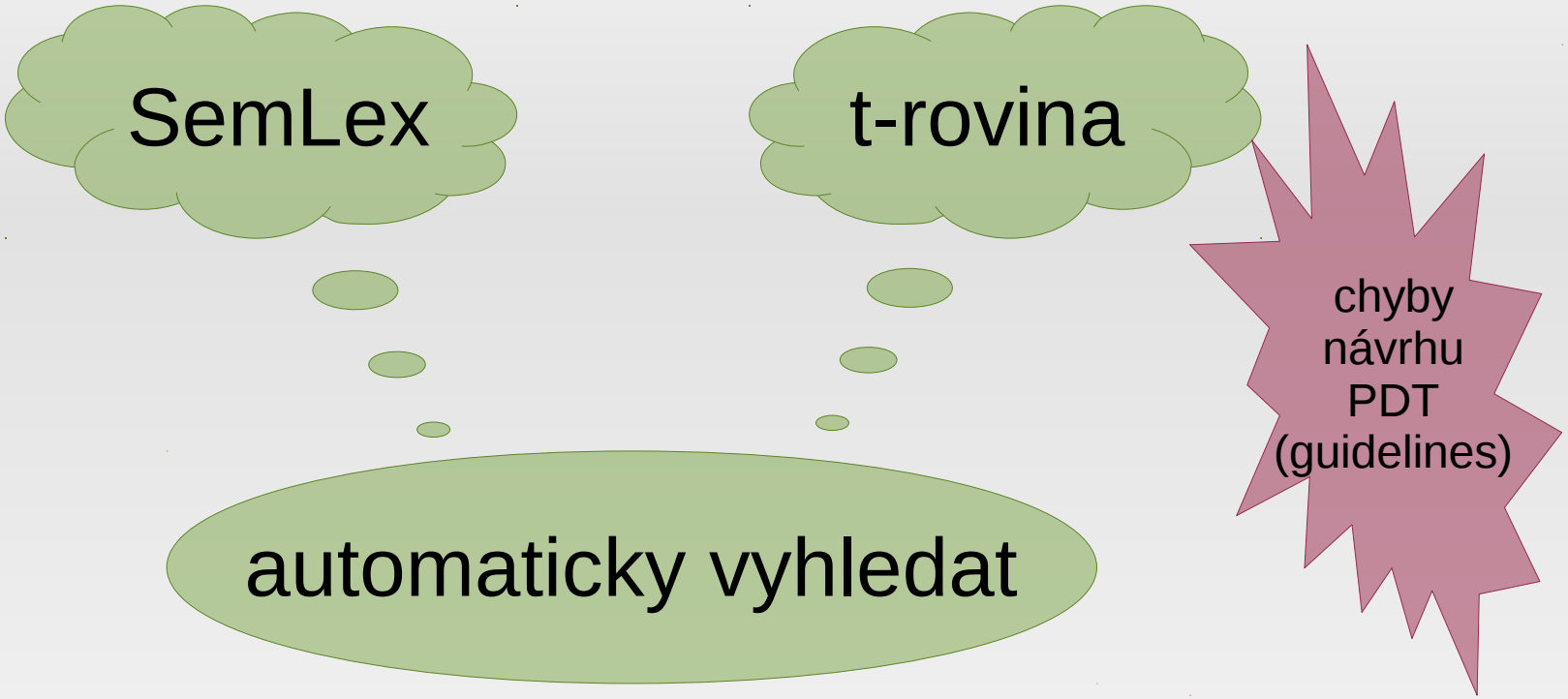
Automatická identifikace



SNADNÉ

...nebýt chyb

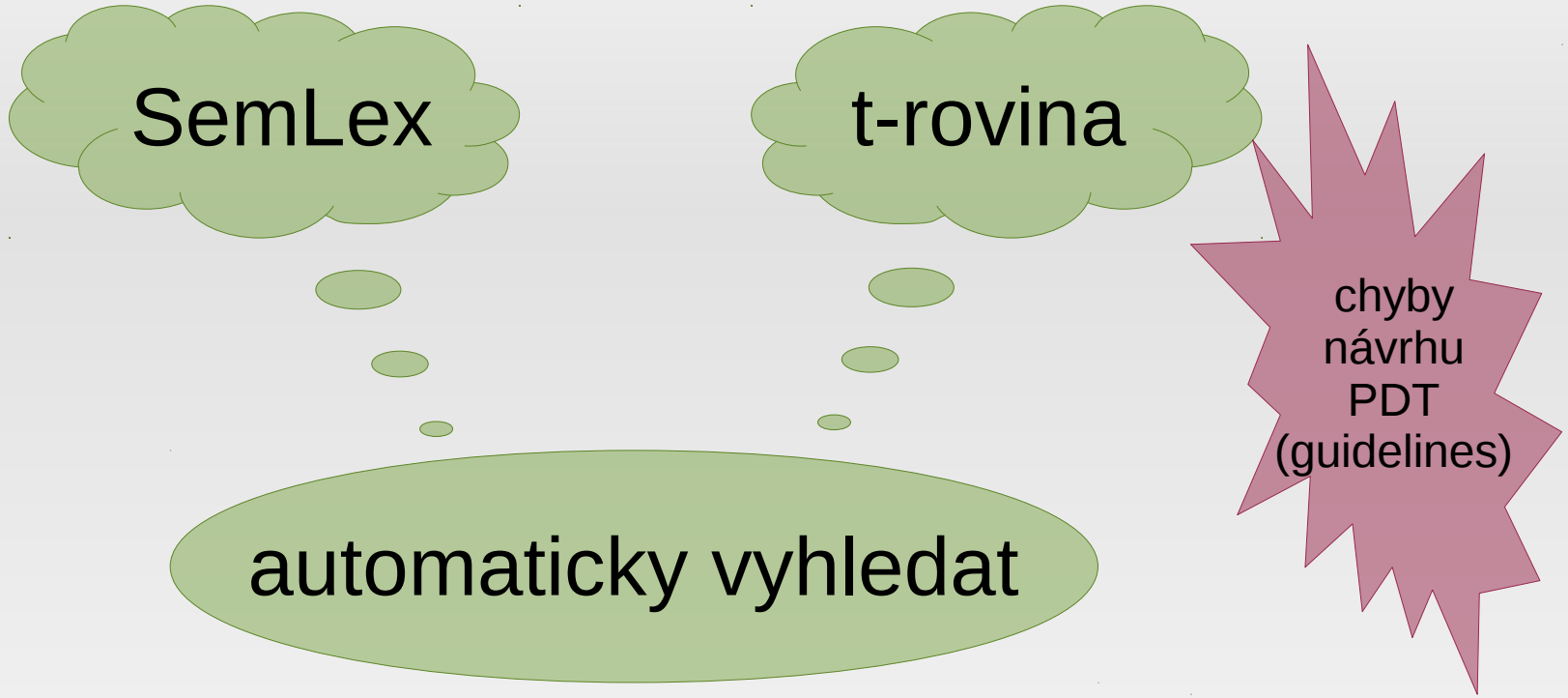
Automatická identifikace



SNADNÉ

...nebýt chyb

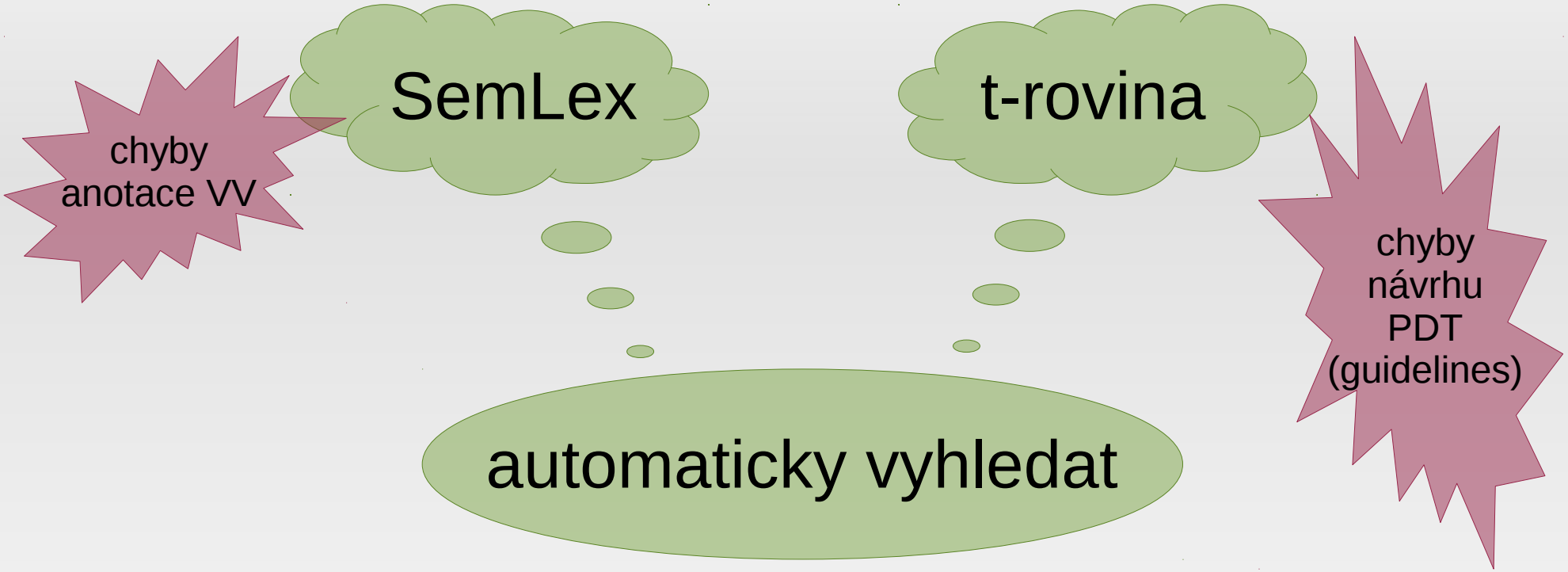
Automatická identifikace



SNADNÉ

...nebýt chyb

Automatická identifikace

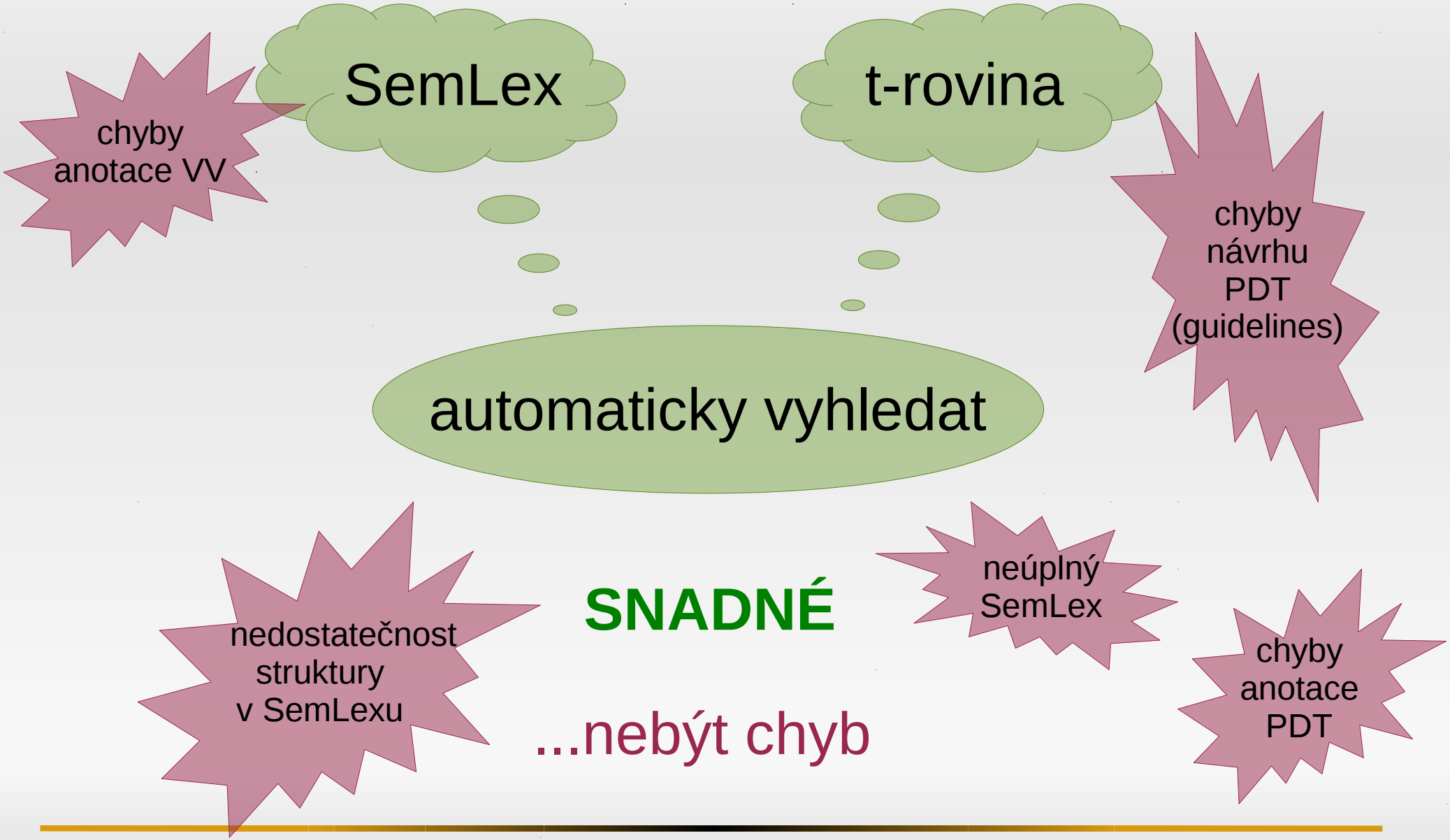


SNADNÉ

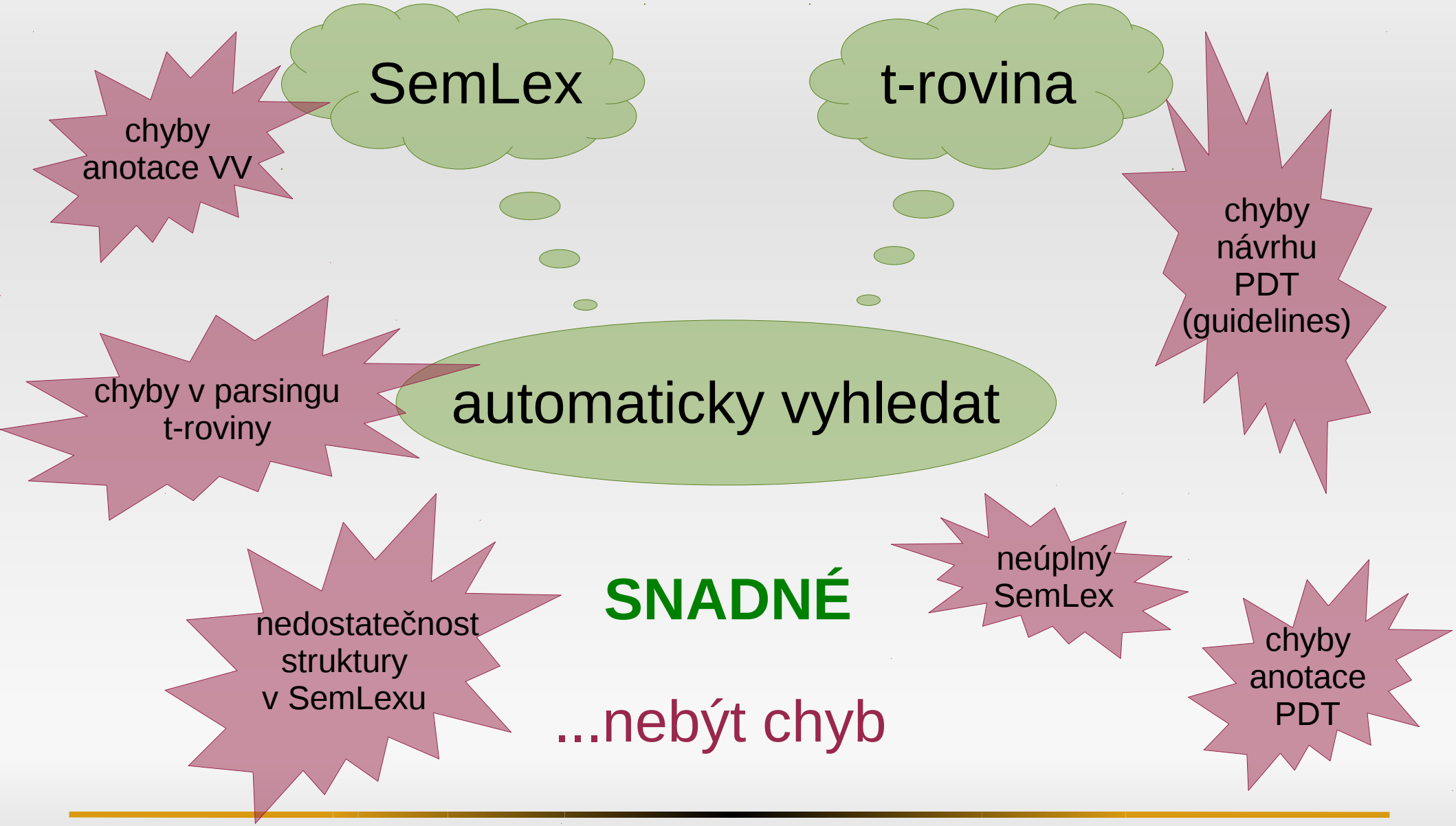
...nebýt chyb

chyby anotace PDT

Automatická identifikace



Automatická identifikace



Podstrom VV nenalezen

- zkratky: *ODS, ministr ŽP*
- přechýlení
  - *mistryně světa, ministryně zahraničí, hlavní hygienička*
- zdrobněliny
  - *rodinný dům/domek, konferenční sál/salonek*
- vid: *zaujmout/zaujímat stanovisko*
- elipsy; reflexivita; PoS; synonyma; přívlastky

Výsledky (zatím žalostné)

Precision / Recall	PDT 2.5	PDT parsované	ČNK parsované
t-rovina	61.99 / **95.95**	63.40 / 86.32	TBD
a-rovina	**66.11** / 88.67	66.09 / 81.96	TBD
m-rovina *(win=3)*	62.65 / 90.50	62.73 / 89.80	TBD

- nedostatečná evaluace na „trénovacích“ datech
- mnoho chyb v datech
- mnoho možných zlepšení pro slovník a hledání
- nezdá se zatím, že by t-rovina porazila a-rovinu, nebo alespoň m-rovinu...

Budoucnost – propojení SemLexu

- „Podobné“ VV jsou i ve slovníku
  - někdy jsou synonymní, či antonymní – jindy zkrátka jen odkazují na tentýž kratší VV
  - nějak je zachytit musíme
  - snaha propojit související pojmy, relacemi (synonymie, hyponymie, ale nejen jimi)
  - případné uznané „duplicity“ promazat
- příklad (násobný)
  - druhá vlna kup**o**nové privatizace, 2. vlna kup**ó**nové privatizace
    druhá vlna KP, druhá vlna privatizace
    druhá vlna

..... může být až 12 variant

Budoucnost – propojení SemLexu

- další příklady (neuspořádané)
  - (základní) umělecká škola
  - (trvale/dlouhodobě) udržitelný rozvoj; (šitý/ušít) na míru
  - osoba/pracovník se změněnou pracovní schopností / změněná prac. sch.
  - ((první) náměstek) ministra zahraničí/zahraničních věcí
  - loutkové divadlo/divadelnictví/představení
  - (deficitní/vyrovnaný/přebytkový) státní rozpočet;
    zákon o (vyr.) st. rozp.;
    (deficit/vyrovnanost/schodek/návrh/výdaje) státního rozpočtu;
    dotace ze státního rozpočtu
  - věci veřejné / věc veřejná / veřejná věc; pro i/a proti

Poděkování Silvii

- Silvie, děkuju. :-)

...a samozřejmě všem za pozornost.

Anotace VV na t-rovině – motivace

- je to blízké anotaci významu, patří to sem
- snazší odstínění povrchových variant (slovosled, nesouvislé/přerušené výrazy)
- tektogramatický podstrom tvořící VV by měl (ideálně) být shodný pro všechny výskyty
- jsou zde doplněné uzly

  popáleniny 3. *a 4.* ***stupně*** *na 40 procentech povrchu těla*

  ...a bez Sarajeva by nebyla žádná první ***světová válka****. A bez první možná ani* ***druhá****.* [PDT 2.0, m-rovina]

  Naučili jsme je ***zobat*** *raději* ***z*** *naší než* ***z*** *jiné* ***ruky****.* [PDT 2.0, m-rovina]

Anotace VV na t-rovině – nevýhody

- vyhledávání pomocí t-podstromu není všemocné:

  *Leonardo **dal** svým gólem signál k výhře nad Nagojou a svůj první **gól** v zemi vycházejícího slunce vstřelil Němec Buchwald.* [PDT 2.0]

- anotace obsahuje pouze t-uzly tvořící VV
  - nevýhoda: aux-uzly nejsou nikdy součástí

    *hodit flintu **do** žita*

    *zkusit **do** třetice*

    *zkoušet **do** soudného dne*

    *investovat **do** kotované emise*
  - lze doplnit automaticky ze slovníku

Tektogramatický podstrom

nevýhody: