S frazémy si hrát na nervy je balzám

Víceslovné výrazy v PDT dnes a zítra

Projekt Lexemann Pavla Straňáka

Workshop grantu Explicitní popis jazyka
a anotovaná data se zřetelem na češtinu

12. dubna 2012

Eduard Bejček

S frazémy si hrát na nervy je balzám

Osnova

- Současný stav
  - PDT 2.0
  - PDT 2.5
- Pravidla anotace víceslovných výrazů v PDT
- Tektogramatická rovina
- Problémové a nečekané výskyty

  aneb proč je nadpis referátu ironický
- Budoucnost

PDT 2.0

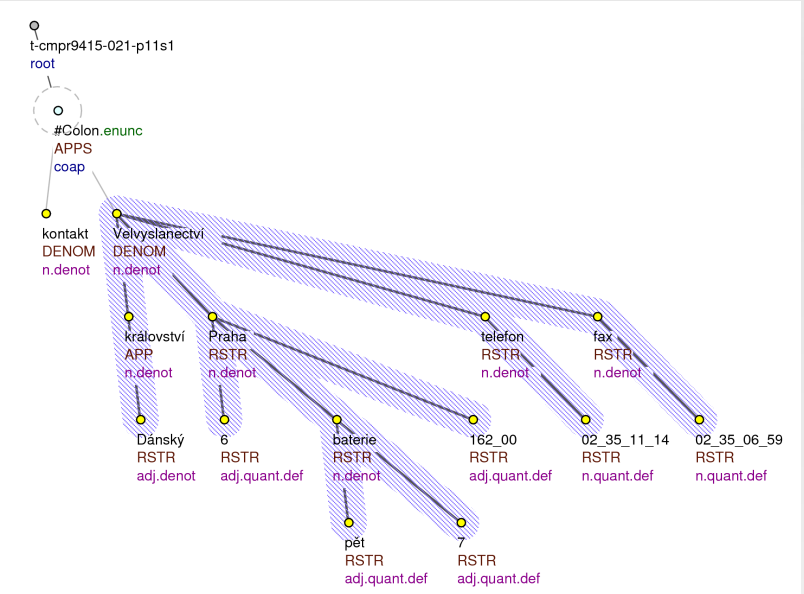
- PDT-VALLEX
- CPHR – složený predikát
- DPHR – (ne)slovesná frazeologická spojení
  - nepatří sem např. *horký brambor* (skloňuje se)
- ID – identifikační výraz (*#Idph*)
  - bez všech výrazů se skloňovaným řídícím členem
- FPHR – cizojazyčná fráze (*#Forn*)
- is_name_of_person

PDT 2.5

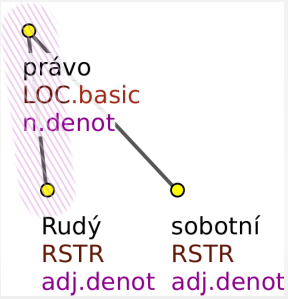
- přibyla ruční anotace VV
  - důsledně na celé t-rovině
  - víceslovné frazémy, idiomy, ustrnulá spojení, ...
  - víceslovné pojmenované entity
- dva druhy zobrazení VV
  - **expandované** (tvar stromu jako dosud, navíc orámovaný podstrom representující VV)
  - **kolabované** (celý VV representován jediným uzlem; vztahy uvnitř jsou skryty)

PDT 2.5 – kolabované uzly

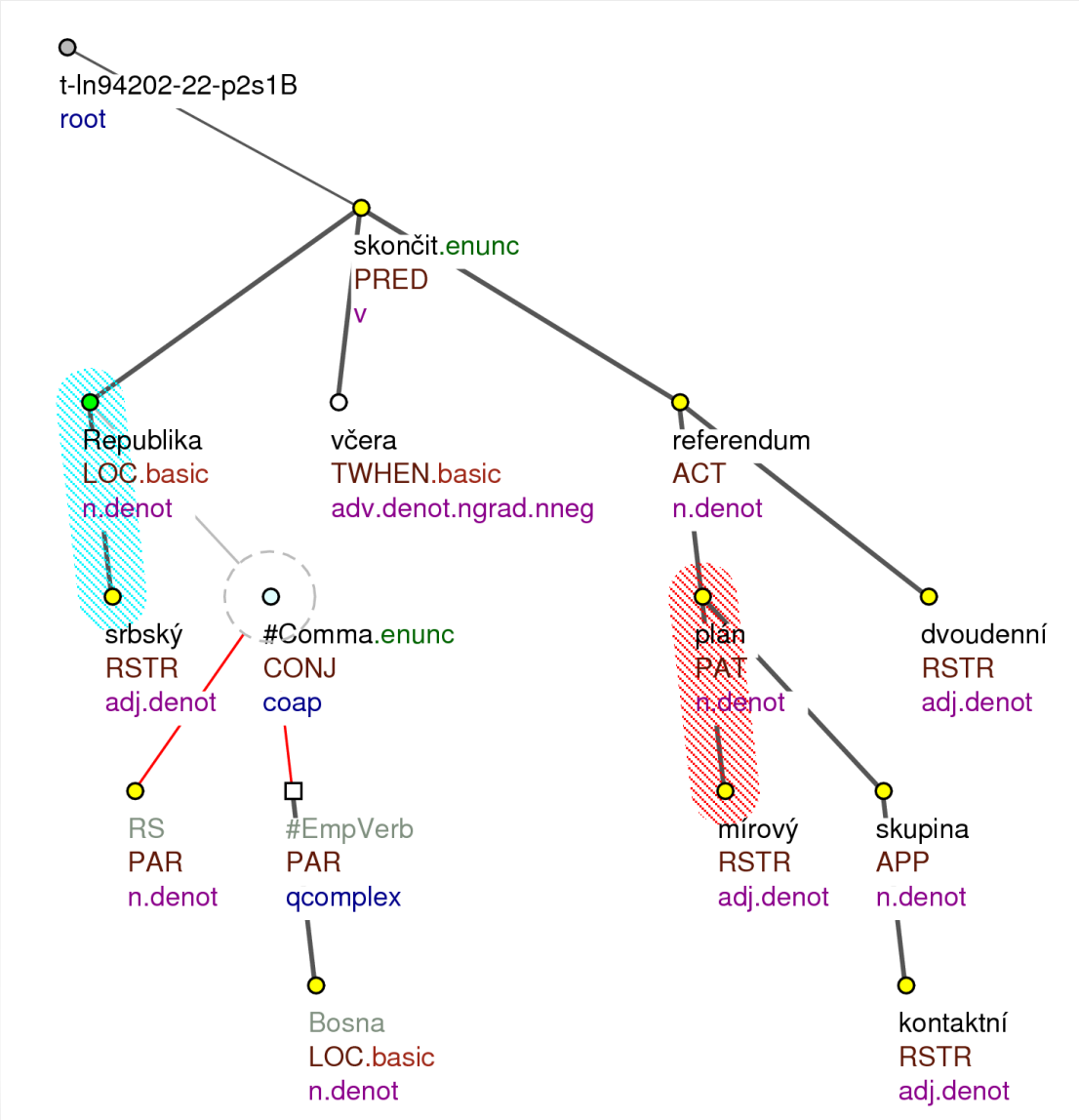
- vztahy uvnitř VV
  - často nejsou závislostní (nebo je význam závislosti jiný než ve zbytku stromu)
  - mnohdy jsou arbitrární (guidelines)
- *„sobotní Rudé právo“*
  - nerozlišitelné přívlastky
- jedno společné „t-lemma“
  - *„olympijské hry“*, ne *„olympijský hra“*



Kontakt: Velvyslanectví Dánského království,
U páté baterie 7, 162 00 Praha 6,
tel.: (02) 35 11 14, FAX: (02) 35 06 59.

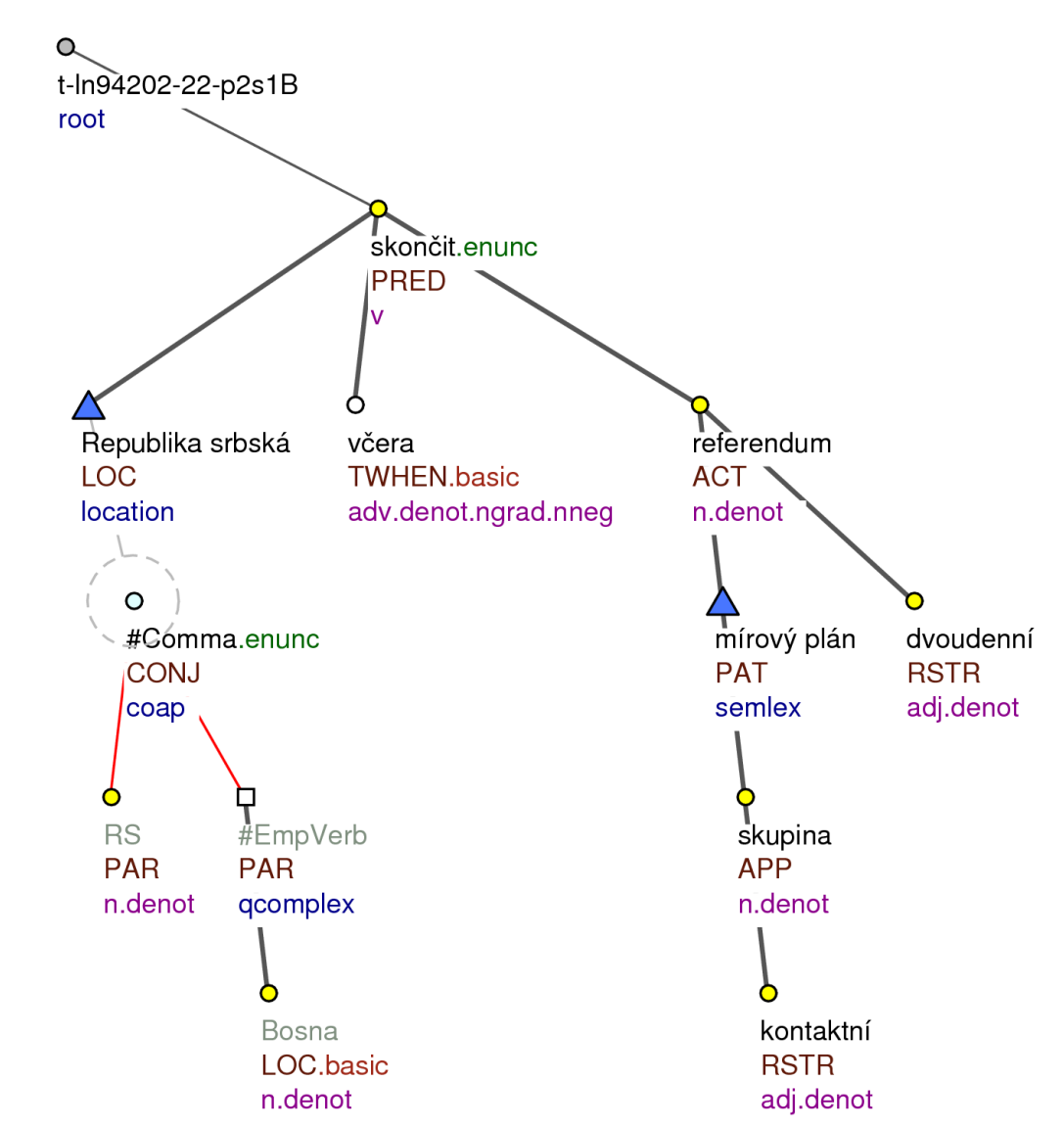


PDT 2.5 – screenshot



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.

S frazémy si hrát na nervy je balzám

Osnova

- Současný stav
  - PDT 2.0
  - PDT 2.5
- Pravidla anotace víceslovných výrazů v PDT
- Tektogramatická rovina
- Problémové a nečekané výskyty
- Budoucnost

Anotace VV

- Jen víceslovné
  - tedy ne „Klaus“, ani „ODS“
  - vyřešení VV ⇒ snazší anotace zbytku
- Dvě skupiny výrazů anotovány současně
  - ustálená spojení, lexémy
    - slovník SemLex (není součástí PDT 2.5)
    - každý VV v textu anotován odkazem do slovníku
  - pojmenované entity
    - osoba, místo, instituce, objekt, adresa, čas, číslo (tedy např. rozsah), cizí výraz, bibliografický údaj
    - žádná další kategorizace, dokonce ani zanoření

Anotace VV – pojmenované entity

- dělení zhruba podle Magdy Ševčíkové (říjen 2005)
- ve většině případů bez apelativ
  - je to problém: „*Mgr. Novotný*“ vs. „*president Obama*“ vs. „*předseda vlády Petr Nečas*“

Anotace VV – typy NE

- person: *profesor* ***P. Novák*** (lze jméno vynechat)
  - též jména zvířat
- institution: ***Dental, s.r.o***; ***galerie Václava Špály***
  - též veletrhy a soutěže
- location: ***New Haven, Connecticut***; *jdu do* ***galerie Václava Špály***; *ČNB* ***Na Příkopech***
- object: ostatní jména (kniha, festival, zákon, zboží, ...)
  - ***Intel Pentium Pro***; ***americký dolar***; ***oxid uhličitý***; ***ČNB*** *Na Příkopech*; ***př. K.***; ***Kč/hod.***

Anotace VV – typy NE (pokrač.)

- address: musí obsahovat ulici a/nebo doplňující nemístní údaj (telefon, PSČ, fax, e-mail, …)
  - ***Velvyslanectví Dánského království, U páté baterie 7, 162 00 Praha 6, tel.: (02) 35 11 14***
  - *Petr Novák, Ústav formální a aplikované lingvistiky*
- biblio: strukturovaný bibliografický údaj
- number: množství: ***100 – 200** metrů čtverečních*; ***od 10 do 18** let*; ***mezi** sedadly **30 – 40***; ***pět a půl procen.***
- time: „kdy“, nikoli „kolik“: *ve* ***20 hodin 15 minut***; ***únor 2002***
- foreign: nepatří-li jinam: ***ad hoc***; *The Jungle Book*

Anotace VV – SemLex

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

Anotace VV – SemLex vs. NE

- Hranice mezi NE a vložením do SemLexu není ostrá.
  - *„ministerstvo pro místní rozvoj České republiky“*
  - *„ministerstvo pro místní rozvoj“*
  - *„krizové centrum pro děti a mládež“*
  - *„životní pojištění“*
- NE lze také vložit do slovníku

S frazémy si hrát na nervy je balzám
Osnova

- Současný stav
  - PDT 2.0
  - PDT 2.5
- Pravidla anotace víceslovných výrazů v PDT
- Tektogramatická rovina
- Problémové a nečekané výskyty
- Budoucnost

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]

Hledání na morfologické rovině

- chybná automatická analýza povrchovou metodou
  - i s očekávanou změnou slovosledu, tvarosloví, prokládání výrazu jinými slovy:

    *Je to balzám **na nervy hrát** s Jenseny. (...) Pierceová po boku Jensenů značně pookřála.* [PDT 2.0, včetně chybějící čárky]

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*

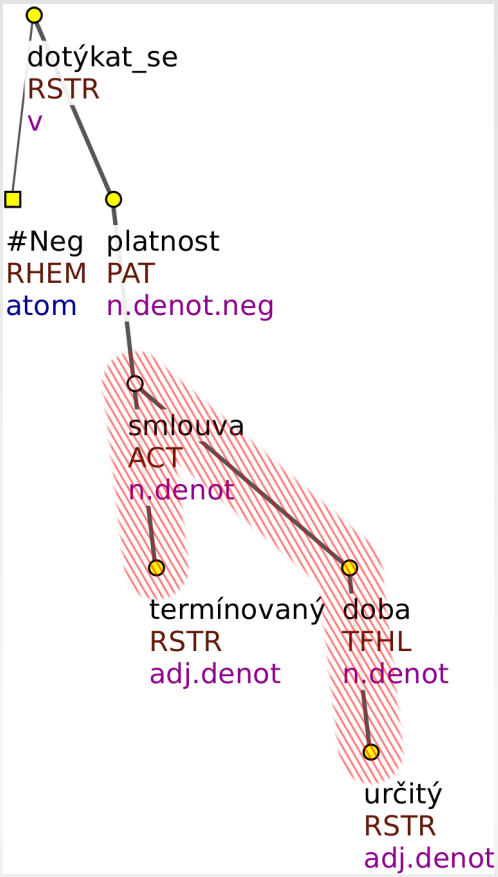
    *zkusit **do** třetice*

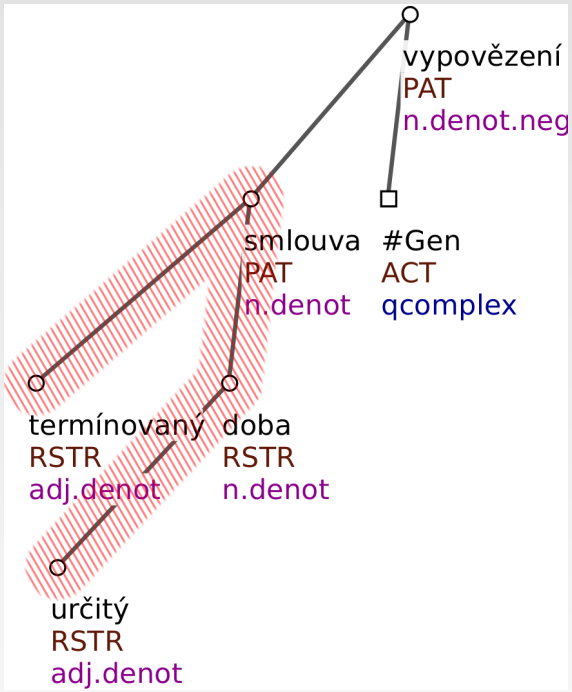
    *zkoušet **do** soudného dne*

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

Tektogramatický podstrom

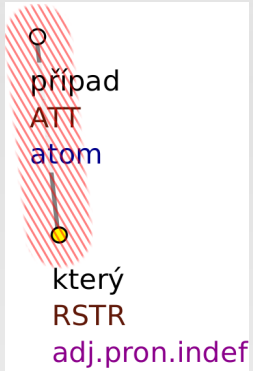
- efektivní rodič
- ideálně je shodný pro všechny instance

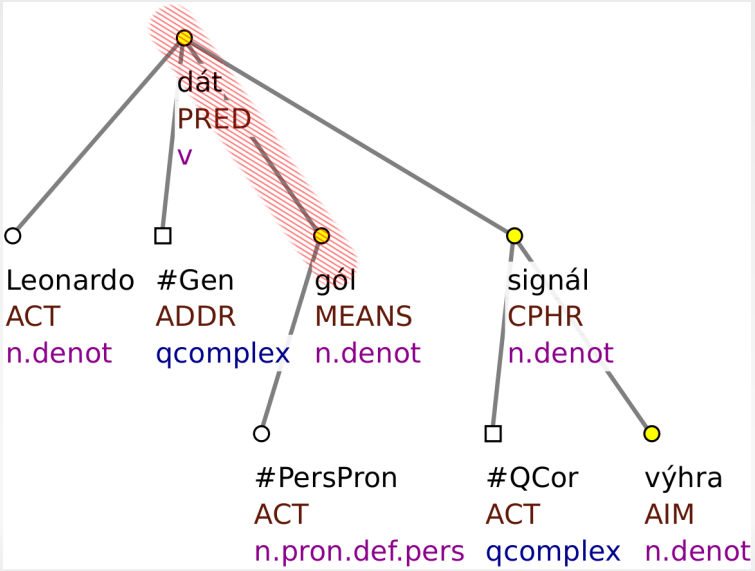




...

nevýhody:



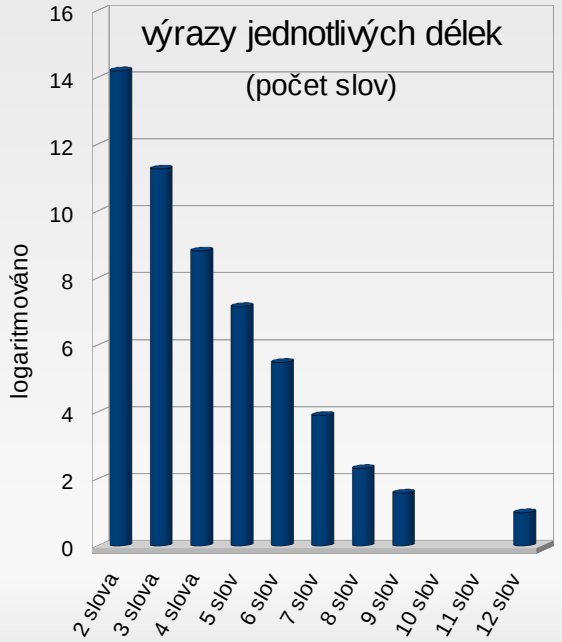


Anotace VV – statistiky v PDT 2.5

- 48.401 VV nalezeno v PDT
  - 45 % tvoří NE
- 8860 položek SemLexu
- 8545 použito v PDT 2.5
- dvojslovné nejčastější
  - 7,5x častější než tříslovné

114 ministr zahraničí
106 akciová společnost
86 státní rozpočet
80 politická strana
78 životní prostředí
74 generální ředitel
70 valná hromada
67 tisková konference
67 ústavní zákon
65 v podstatě
65 mimo jiné

7801 person
5533 institution
4023 number
3226 object
3001 time
2091 location
157 foreign
146 address
40 biblio



S frazémy si hrát na nervy je balzám
Osnova

- Současný stav
  - PDT 2.0
  - PDT 2.5
- Pravidla anotace víceslovných výrazů v PDT
- Tektogramatická rovina
- Problémové a nečekané výskyty
- Budoucnost

Tektogramatický strom – neshody

- zkratky: *ODS, ministr ŽP*
- přechýlení
  - *mistryně světa, ministryně zahraničí, hlavní hygienička*
- zdrobněliny
  - *rodinný dům/domek, konferenční sál/salonek*
- vid: *zaujmout/zaujímat stanovisko*
- *být/končit na mrtvém bodě, občanský zákon/zákoník, cenová hladina/hladina cen, ...*

Objívky

- t-uzly representující VV nemusí tvořit souvislý podstrom
  - obvykle však jde o patologické případy, nebo „chyby“ anotace PDT
  - *k.o.* (visí na #Forn)
  - *zakopaný pes*
  - *být* ***bojovník***.ACT ***ve zbrani***.MANN
  - *bylo to asi tak* ***padesát***.MANN ***na padesát***.MANN
- VV mohou stále ještě být homonymní
  - *přímá volba*

Objívky

- Parafráze, exploatace
  - autoři k dosažení svého efektu pokřivují ustálená spojení
  - vyznam se v čase mění, posouvá

  - *Zloději nechodí po horách, ale po domácnostech.*
  - *ne Rudé, ale Šedé právo*
  - *Sarajevo* – (v českém kontextu často) nejde o místo, ale o událost
  - *Sarajevský atentát* – nový význam

Objívky

- některé frazémy by bylo lepší zachytit pomocí wildcardů
  - *stát {v popředí, ve stínu, v pozadí, na výsluní}*
  - *dostat [0-9]+ {měsíců,roků} {natvrdo,podmíněně}*
  - *drží {první,druhou,třetí,...,předposlední,...} příčku*

Objívky

- některé NE mohou být až nečekaně dlouhé:
  - *The Black Box Summer Festival of Czech 20th Century Plays – Vašek Káňa: Karhan' s Men*
- „slabší“ třídu frazémů lze přerušit užitím spony
  - *škola chce mít* ***třídy smíšené***
  - *věří, že* ***kurs*** *bude do konce roku* ***volně plovoucí***
  - *?****dům*** *na konci ulice je* ***rodinný***
  - *****pomoc****, kterou jsem zraněné poskytl, byla* ***první***

Budoucnost

- kostlivci ve skříni:
  - is_name_of_person
  - automatická kontrola anotátorských „opomenutí“
- kategorizace pojmenovaných entit
  - tabulka jako struktura adresy
- vnořené pojmenované entity
  - vnoření odkazem ve slovníku
    - *náměstí Jiřího z Poděbrad 34*
- jednoslovné pojmenované entity

#1342: Poděbrady (location)
#32959: Jiří z #1342 (person)
#45323: náměstí #32959 (location)

město	–
ulice	#45323
č.p.	–
č.o.	34
...	...

(address)