Predstavljanje polileksičkih jedinica u bankama drveta zavisnosti

PARSEME COST akcija

Vesna Pajić i Staša Vujičić Stanković
Društvo za jezičke resurse i tehnologije

Sadržaj izlaganja

- O COST akcijama i PARSEME akciji
- Polileksičke jedinice (*Multiword Units* – MWE)
- Predstavljanje zimske škole o polileksičkim jedinicama i njihovoj obradi
  - Predavači
  - Predavanje
  - Teme, diskusije, problemi
- Zaključak

COST projekat

COST

- 1971. godine je započeo kao Evropski projekat
- Osnovna namena – podrška međunarodnoj saradnji istraživača, inženjera i akademaca širom Evrope
- Podrška i dopuna aktivnostima u okviru drugih EU programa
- 2013. godine je ustanovljena COST asocijacija, koja nastavlja da vodi sve tekuće COST aktivnosti.
- U ovaj projekat je trenutno uključeno 36 evropskih zemalja i 1 država-saradnik što omogućava istraživačima iz ovih zemalja da učestvuju u COST akcijama

Osnovni principi

- **Internacionalnost**
- **Propagiranje ravnopravnosti**
- ***Inclusiveness*** (uključivanje istraživača iz zemalja koje su prepoznate kao zemlje sa niskim istraživačkim kapacitetima, jedna od njih je i Srbija)
- **Otvorenost i interdisciplinarnost**
- **Podrška mladim istraživačima** (ESR – *Early Stage Researchers* – Ph. D. studenti ili doktori nauka koji su doktorirali pre manje od 8 godina)

Kako radi COST?

- Ne finansira istraživanja, već samo aktivnosti vezane za međunarodnu saradnju (troškovi putovanja i boravka, troškovi za lokalnu organizaciju događaja, troškovi publikacija...)
- Mreže istraživača, inženjera i akademaca se okupljaju oko COST akcija, koje su glavni nosioci aktivnosti

Aktivnosti:

sastanci, radionice, konferencije, treninzi,
kratkotrajne naučne posete (*STSMs*) i
aktivnosti vezane za širenje znanja (*dissemination*)

Struktura COST akcija

- Svaka akcija ima svoju temu oko koje se odvijaju sve aktivnosti, jasno definisane ciljeve i rezultate koje treba ispuniti po završetku akcije
- Sve odluke o akciji donosi upravljačko telo akcije (*MC – Management Commitee*) sastavljeno od po najviše dva predstavnika za svaku zemlju članicu i njihovih zamenika
- Rad se odvija kroz radne grupe
- Period trajanja svake COST akcije je 4 godine
- Zvanični veb sajt svake akcije sadrži sve potrebne informacije

COST Actions

PARSEME COST akcija

PARSEME

- Akcija iz domena:
  Informacione i komunikacione tehnologije, broj IC1207
- Zvaničan sajt akcije:
  http://www.parseme.eu
- Trajanje:
  od 08. marta 2013. do 07. marta 2017. godine
- Predsedavajući akcije:
  Agata Savary (FR) i Adam Przepiorkowski (PL)

PARSEME

- PARSEME (PARSing and Multi-word Expressions):
  Towards linguistic precision and computational efficiency in natural language processing

- Osnovu istraživanja u okviru ove akcije čine ***polileksičke jedinice***, njihovo predstavljanje u jezičkim resursima, kao i razvoj tehnologija i softverskih alata za njihovo detektovanje u tekstu, razumevanje, prevođenje...

Polileksičke jedinice

- Radna definicija:

Polileksičke jedinice su izrazi koji se sastoje iz više od jedne reči i moraju da budu uskladišteni u rečnike zbog svog osobenog ponašanja u smislu pravopisa, morfologije, sintakse ili semantike.

Polileksičke jedinice

- Polileksičke jedinice čine preko 40% teksta
- Njihova kompleksnost se ogleda na više nivoa:
  - značenje se ne prenosi direktno kompozicijom značenja njihovih sastavnih delova (*„pritegnuti kaiš“*),
  - imaju određene sintaksne varijacije (*pritegnuo je/pritegnula je/pritegnuti kaiš*),
  - mogu biti leksički ili sintaksno osobene (*vojna tajna*, *glava porodice, biti „široke ruke“*)

PARSEME

Polileksičke jedinice – otvorena pitanja

- Još uvek nisu potpuno objašnjene i shvaćene
- Teške su za detektovanje, razumevanje, prevođenje...
- Nisu dovoljno zastupljene u jezičkim resursima i softverskim alatima

Koje izraze upisati u elektronski resurs?

Bela vrana? *Pritegnuti kaiš*? *Vojna tajna*?

Koji format odabrati? Kako ih obeležiti?

PARSEME

PARSEME

Radne grupe u okviru akcije:

- RG1: Leksičko-gramatičko sučelje
- RG2: Tehnike parsiranja polileksičkih jedinica
- RG3: Statistička, hibridna i višejezična obrada polileksičkih jedinica
- RG4: Anotiranje polileksičkih jedinica u bankama drveta

RG4: Anotiranje polileksičkih jedinica u bankama drveta

- Metodologije izgradnje banaka drveta
- Optimalna upotrebljivost polileksičkih jedinica u parsiranju

Kao jedna od aktivnosti ove radne grupe održana je prva škola (trening) u okviru PARSEME akcije.

Zimska škola o polileksičkim jedinicama i njihovoj obradi

Zimska škola u Pragu

- Prva zimska škola održana je od 19. do 23. januara 2015.
- Zvanična veb prezentacija sa slajdovima i informacijama https://ufal.mff.cuni.cz/events/parseme-1st-training-school
- Mesto održavanja: Institut za formalnu i primenjenu lingvistiku, Karlov Univerzitet u Pragu
- Trajanje: 5 radnih dana po 4 sekcije (predavanja i praktične vežbe na računarima)

Treebanks (Banke drveta)

- Osnovna struktura kojom smo se bavili u Pragu
- *Treebank* (Banka drveta ???) je parsirani tekstualni korpus u kome je obeležena sintaksna ili semantička struktura rečenice.
- Obično se kreiraju nad korpusima u kojima su već označene vrste reči, a dodatno je moguće banke drveta dopuniti i semantičkim ili drugim lingvističkim informacijama.
- Mogu biti kreirane ručno, poluautomatski ili automatski.

Banke drveta

Dve glavne grupe

- One koje anotiraju strukturu fraze (takve su *PennTreebank* ili *ICE-GB*)
- One koje anotiraju strukturu zavisnosti (npr. *Prague Dependency Treebank* ili *Quranic Arabic Dependency Treebank*).

U okviru PARSEME radne grupe 4 pokrenuta je akcija prikupljanja informacija o postojećim bankama drveta, vrstama polileksičkih jedinica i načinu na koje su one predstavljene u njima.

(http://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme)

Treebank	Language	Nominal MWEs			Verbal MWEs				Prepositional MWEs	Adjectival MWEs	MWEs of other categories	Proverbs
		Named entities	NN compounds	Other nominal MWEs	Phrasal verbs	Light verb constructions	VP idioms	Other verbal MWEs				
The INESS Norwegian Treebank	Norwegian	YES	N/A	YES	YES	NO?	YES?	NO?	YES	YES	NO?	NO
The Lassy Small Treebank	Dutch	YES	YES	YES	YES	COMP	COMP	NO	YES	NO	NO	NO
DeepBank	English	YES?	YES	YES?	YES	YES?	YES?	YES?	YES?	YES?	YES?	YES?
META-NORD Sofie Swedish Treebank	Swedish	YES	NO?	YES?	YES?	YES?	YES?	YES?	YES?	YES?	YES?	YES?
BulTreeBank	Bulgarian	YES	N/A	YES	N/A	COMP	COMP	NO	YES	YES	YES	COMP
The Prague Dependency Treebank	Czech	YES	YES	YES	yes	YES	yes	yes	NO	yes	yes	YES
The French Treebank	French	YES	YES	YES	N/A	NO	YES	NO	YES	YES	YES	NO
The Cintil Portuguese Treebanks	Portuguese	YES	COMP	N/A	N/A	COMP	N/A	N/A	YES	N/A	YES	COMP

Predavači i teme

Manfred Sailer (Frankfurt) Shuly Wintner (Haifa)	Lingvističke osobine polileksičkih jedinica. Polileksičke jedinice u lingvističkoj teoriji (engleski/nemački/francuski). Izazovi u drugim jezicima (hebrejski). Kodiranje i primene.
Dan Flickinger (Stanford)	Uvod u gramatike za opisivanje strukture fraza u obliku „glava – zavisni deo“ i izazovi koji potiču od polileksičkih jedinica.
Joakim Nivre (Uppsala)	Uvod u gramatike za opisivanje zavisnosti između reči u rečenicama i parsiranje. Parsiranje zavisnosti zasnovano na grafovima i prelazima (*transition based*). Polileksičke jedinice u parsiranju zavisnosti- Praktični rad u laboratoriji sa *MaltParserom*.
Jan Hajič Pavel Straňák Jiří Mírovský (Prague)	Banke drveta i polileksičke jedinice.

Praška banka drveta
The Prague Dependency Treebanks - PDT

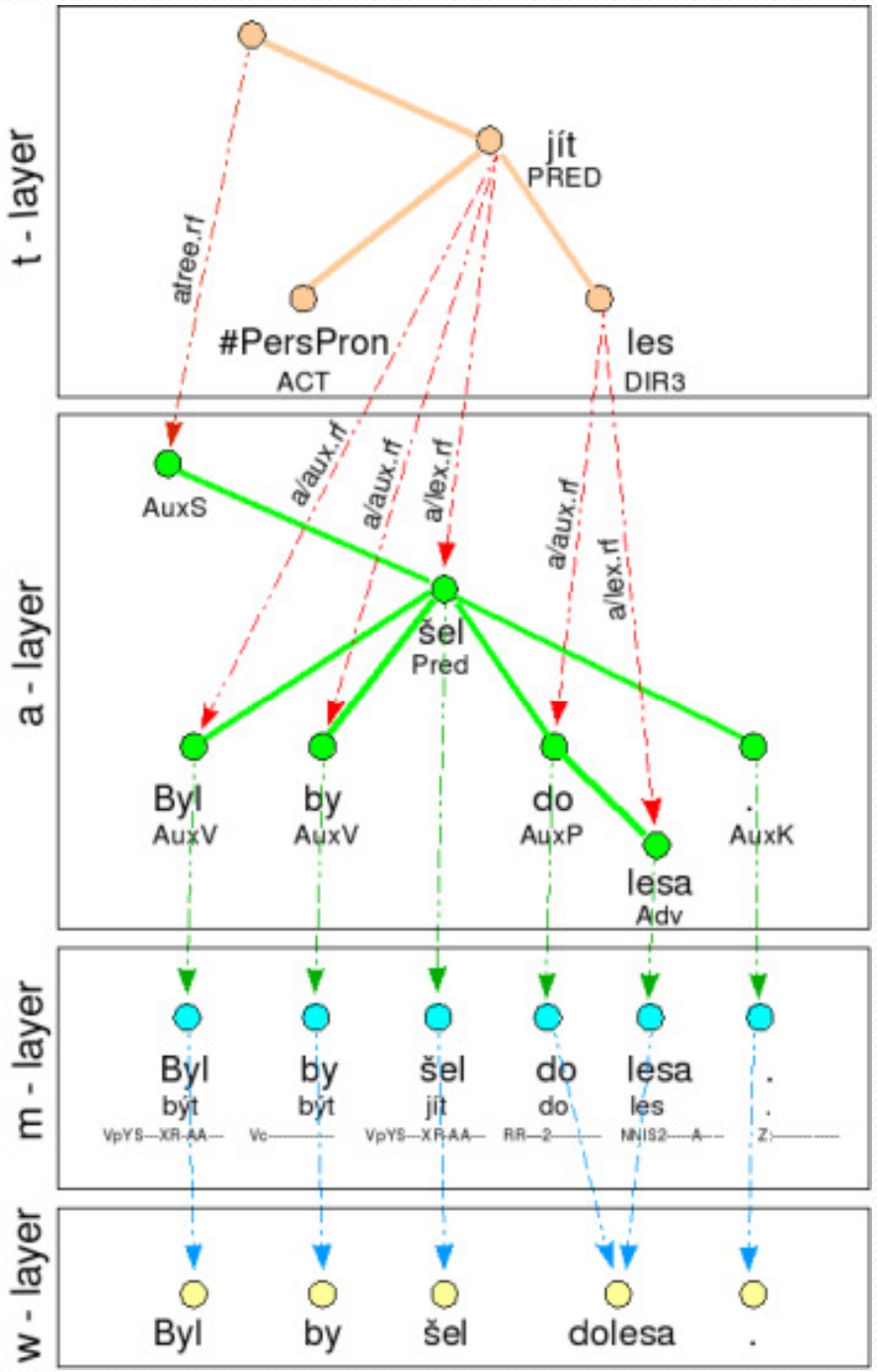
- Prva verzija PDT 1.0 nastala je 2001
- Trenutno aktuelna verzija - PDT 3.0: http://ufal.mff.cuni.cz/pdt3.0
- Jezici: češki, engleski, arapski
- Češki – tekstovi su iz 4 izvora, od toga dve dnevne novine, jedan popularni naučni magazin i jedan ekonomski žurnal (od 1990. do 1995.)
- Osnovna jedinica korpusa (dokument) je članak
- 1.8 miliona tokena (~110,000 rečenica)

PDT - osobine

- Sastoji se od četiri sloja informacija
- Tri sloja su anotirana:
  - Sloj morfema
  - Analitički sloj (sintaksa)
  - Tektogramatički (eng. *tectogrammatical*) sloj: koreference, valentnost, semantika...

- Format:
  - *Prague Markup Language* (zasnovan na XML-u)
  - U novije vreme razvijen i *.treex* format za upotrebu na TreeX platformi (Perl)

Slojevi u PDT



Morfološki sloj anotacije

- U okviru ovog sloja svaka reč biva označena odgovarajućim oznakama, iz ukupno 13 kategorija
- Svakoj reči se pridružuje tačno jedna lema
- Dodeljivanje leme je rađeno ručno, pri čemu je tačna lema birana na osnovu konteksta

Morfološki atributi

Pr.: nejnezajímavějším
„najnezanimljiviji“

Oznake: 13 kategorija

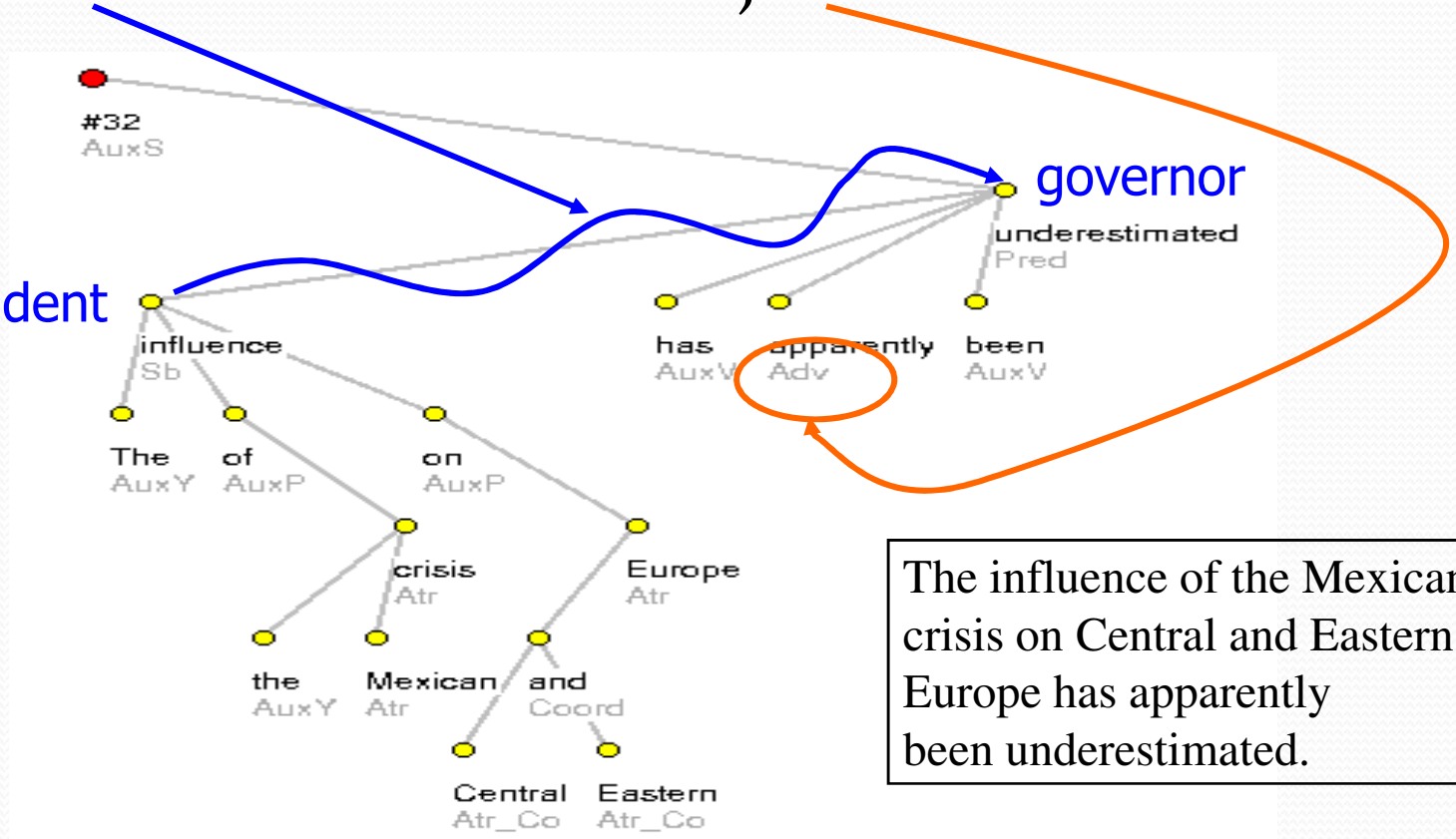
Primer: AAFP3----3N----

Lema:

Books/verb -> book-1, went -> go, to/prep. -> to-1

Analitički sloj anotacije

- Zavisnost+ Analitička funkcija



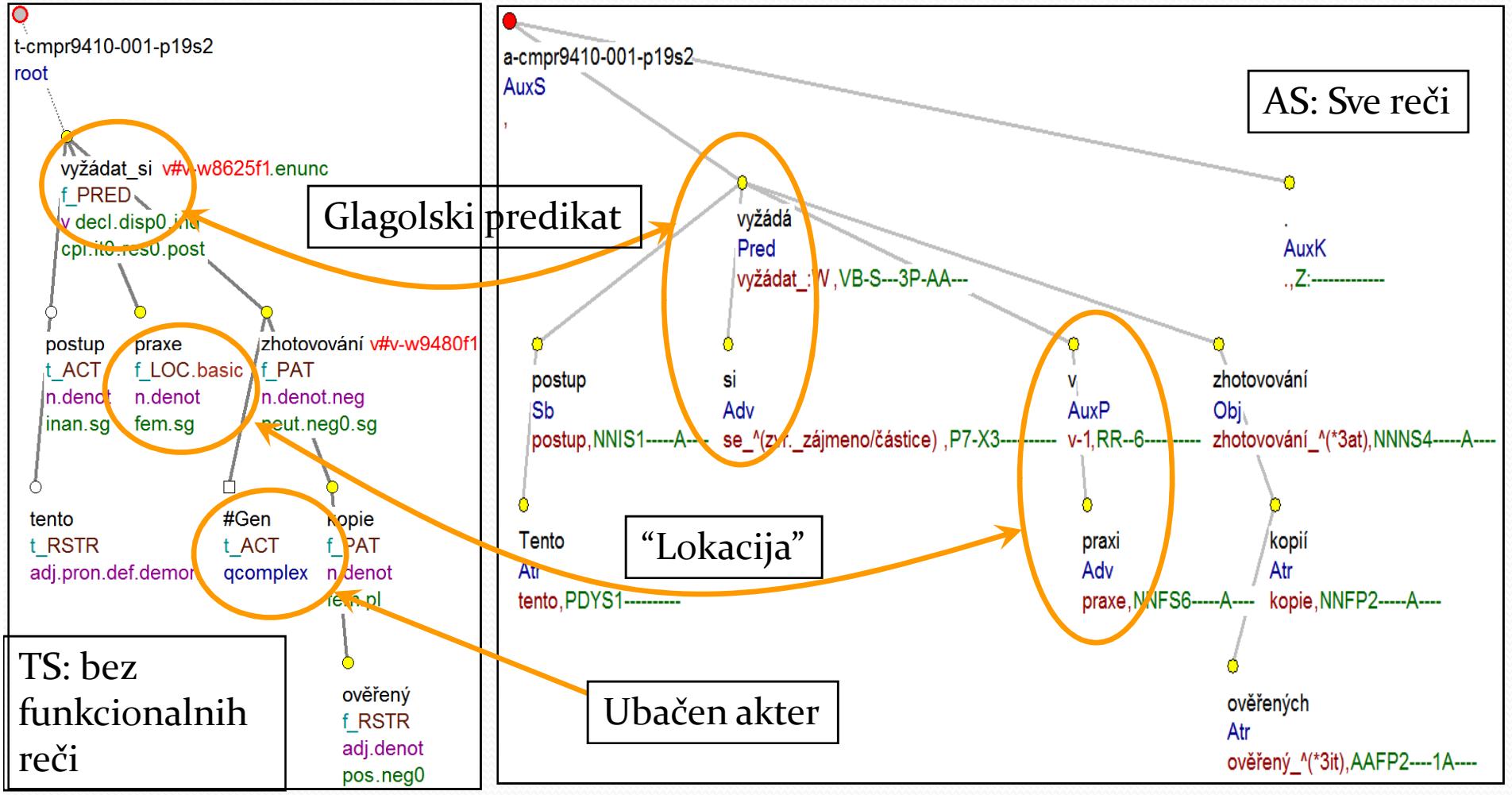
The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated.

Analitičke funkcije

- Osnovne:
  - Pred, Sb, Obj, Adv, Atr, Atv(V), AuxV, Pnom
  - „dupla“ zavisnost: AtrAdv, AtrObj, AtrAtr
- Specijalne (funkcionalne reči, znakovi interpunkcije ...):
  - AuxT, AuxR, AuxO, AuxZ, AuxY
  - predlozi: AuxP, AuxC
  - interpunkcija, grafički elementi: AuxX, AuxS, AuxG, AuxK
- Strukturalne ...
  - elipse i druge: ExD, Coord, Apos

Tektogramatički sloj

- Opisuje dubinsku sintaksu
- Sadrži 5 podslojeva: struktura zavisnosti, reči u fokusu, koreference, diskurs, sve drugo...
- Ukupno: 39 atributa



In practice, that procedure will require making of certified copies.

Struktura tektogramatičkog sloja

- Određeni čvorovi su obrisani
  - pomoćni glagoli, članovi, interpunkcija…
  - neke polileksičke jedinice se spajaju u jednu
- Određeni čvorovi se dodaju
  - na osnovu valentnosti glagola
  - razrešenja elipsi
- Obeležava se detaljno relacija zavisnosti

Funktori tektogramatičkog sloja

- U tektogramatičkom sloju se koristi veliki broj funktora
- Akteri: ACT, PAT, EFF, ADDR, ORIG
- Razni drugi:
  - LOC, DIR1, ...; TWHEN, TTILL,...; RSTR; BEN, ATT, ACMP, INTT, MANN; MAT, APP; ID, DPHR, CPHR, ...
  - Coordination, Rhematizers, Foreign phrases (#Forn),...

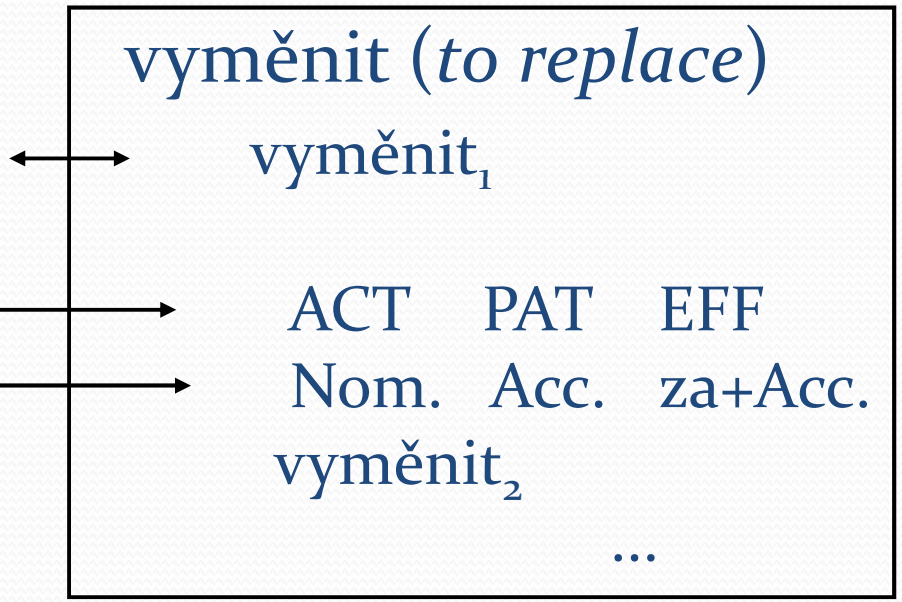
MWEs: DPHR (dependent part of phraseme), CPHR (compound phraseme), FPHR (foreign phrase)

Valentnost u PDT

- Značajna sa stanovišta obeležavanja i prepoznavanja polileksičkih jedinica
- Označava sposobnost reči (najčešće glagola) da za sebe vežu druge reči, da traže dopune (obavezne ili opcione) i da u rečenici otvaraju prazna mesta i regulišu njihovo popunjavanje.
- Specifična je za pojedinačna značenja reči
  - leave: *sb left sth for sb* vs. *sb left from somewhere*
- Najčešće je u vezi sa oblikom reči

Format odrednice u leksikonu valentnosti

- reč(lema)
  - značenje reči 1
    - Okvir valentnosti:
      - $slot_1$ $slot_2$ $slot_3$
    - Sintaksni opis
  - Značenje reči 2
    - …



Prague Markup Language PML

PML

- Zasnovan na XML-u
- Opšti jezika za predstavljanje drvolikih struktura i njihovih povezanosti
- Koristi se za predstavljanje PDT, pri čemu omogućava povezivanje sva 4 sloja kao i povezivanje sa leksikonom valentnosti
- Koristi se i za druge namene i druge banke drveta:
  - Treex sistem
  - Learner corpus CzeSL
  - **Treebanks in PML:** https://lindat.mff.cuni.cz/services/pmltq/

PML alati i biblioteke

Biblioteke:

- Perl: Treex::PML package (CPAN)
- Java biblioteke

Alati:

- Tred - uređivanje drveta, pretraga drveta
- PML Tree Query: https://lindat.mff.cuni.cz/services/pmltq/
- MEd: poravnanja prevoda, audio transkripcije
- Law: alat za morfološku anotaciju
- Feat: slojevita anotacija korpusa za učenje
- Capek: alat za anotaciju namenjen školskoj deci za izučavanje gramatike

Strukture podataka u PML-u

- *atomic* – formatirana niska
- *enumerated type* – zadati skup mogućih vrednosti
- *structure* – skup parova atribut-vrednost
- *list* – uređena ili neuređena lista elemenata istog tipa
- *sequence* – lista koja može da ima elemente različitog tipa

PML šema

```
<?xml version="1.0"?>
<pml_schema version="1.1"
    xmlns="http://ufal.mff.cuni.cz/pdt/pml/schema/">
<description>Example of constituency tree
    annotation</description>
<root name="annotation">
<sequence role="#TREES" content_pattern="meta, nt+">
<element name="meta" type="meta.type"/>
<element name="nt" type="nonterminal.type"/>
</sequence>
</root>
<type name="meta.type">
<structure>
<member name="annotator"><cdata format="any"/></member>
<member name="datetime"><cdata format="any"/></member>
</structure>
</type>
<type name="nonterminal.type">
<container role="#NODE">
<attribute name="label" type="label.type"/>
<sequence role="#CHILDNODES">
<element name="nt" type="nonterminal.type"/>
<element name="form" type="terminal.type"/>
</sequence>
</container>
</type>
<type name="terminal.type">
<container role="#NODE">
<cdata format="any"/>
</container>
</type>
<type name="label.type">
<choice>
<value>S</value>
<value>VP</value>
<value>NP</value>
<!-- etc. -->
</choice>
</type>
</pml_schema>
```

Primer označenog drveta

```
<?xml version="1.0"?>
<annotation xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
<head>
    <schema href="example_schema.xml"/>
</head>
<meta>
    <annotator>John Smith</annotator>
    <datetime>Sun May 1 18:56:55 2005</datetime>
</meta>
<nt label="S">
    <nt label="NP">
            <form>John</form>
    </nt>
    <nt label="VP">
            <form>loves</form>
            <nt label="NP">
                    <form>Mary</form>
            </nt>
    </nt>
</nt>
</annotation>
```

Označavanje polileksičkih jedinica u PDT

- Polileksičke jedinice sadržane u jednoj rečenici su smeštene u atribut nazvan mwes u okviru korenog čvora tektogramatičkog drveta te rečenice.
- Svaka MWE je izlistana u atributu mwes i sadrži ID, osnovnu formu, tip i listu identifikatora t-čvorova koji je sačinjavaju.
- Ukoliko je polileksička jedinica imenovani entitet, navodi se njegov tip:
  - *"lexeme"* – polileksička leksema
  - *"person"* – ime osobe ili životinje
  - *"institution"* – institucija
  - *"location"* – geografska lokacija
  - *"object"* – knjiga, jedinica mere, biološko ime biljke ili životinje
  - *"address"* – adresa
  - *"time"* – izrazi koji označavaju vreme ili datum
  - *"biblio"* – bibliografska referenca
  - *"foreign"* – strani izraz
  - *"number"* – numerička vrednost, obično raspon

PML-TQ Query

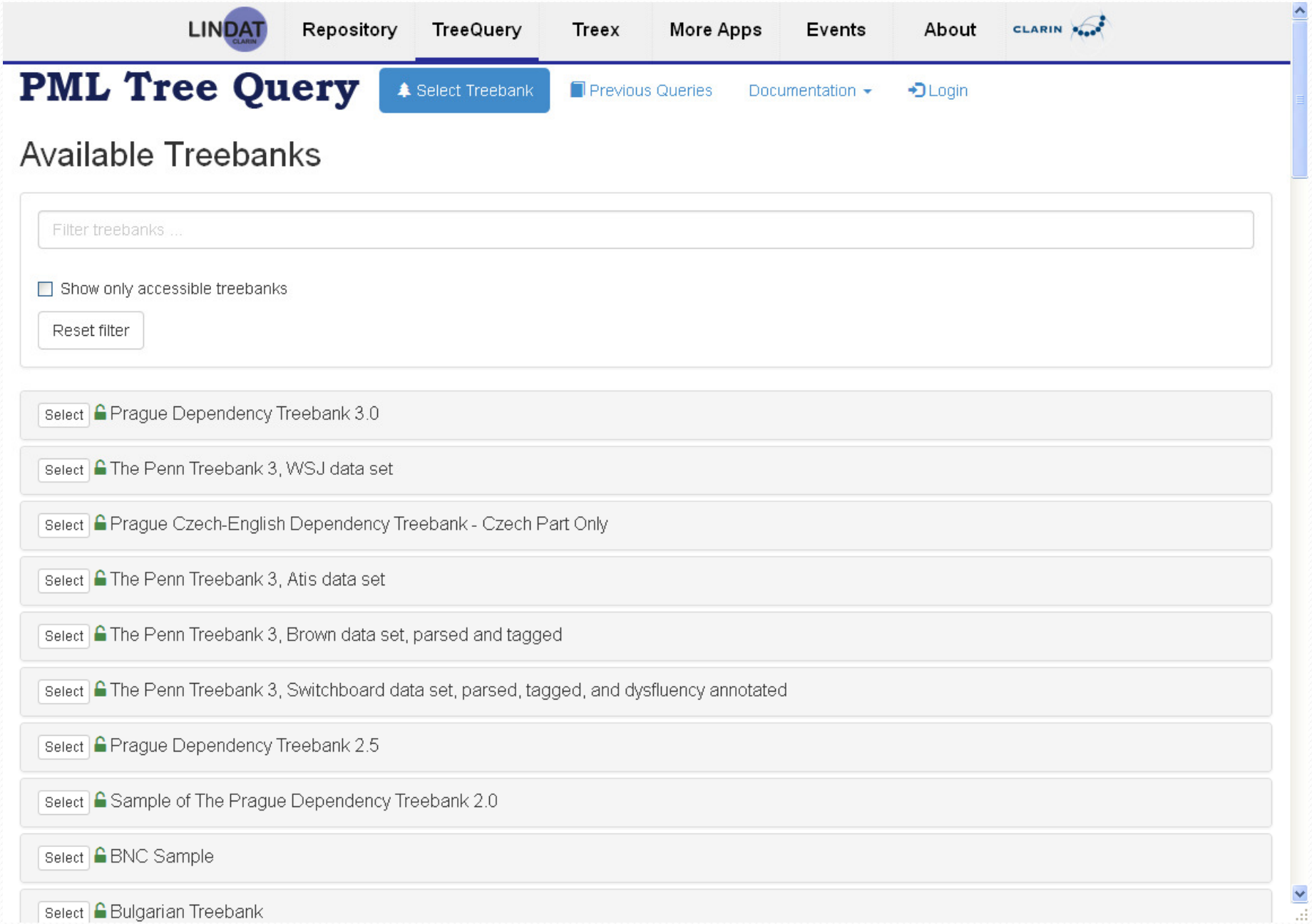
PML-TQ alat za postavljanje upita

Sastoji se iz tri komponente:

- **Upitni jezik** koji podržava upite nad više slojeva i nad kompleksnim strukturama podataka. Uključuje i podjezik za generisanje netrivijalnih statističkih izveštaja
- **GUI** sa alatom za izradu upita, vizuelizaciju rezultata, klijentsku veb aplikaciju i sučelje za rad iz komandne linije
- **Dva softverska procesora** za evaluaciju upita: jedan brz i efikasan koji zahteva da banka drveta bude transformisana u relacionu bazu podataka i drugi sporiji koji radi direktno nad datotekama banke drveta.

Mogućnosti PML-TQ alata

- Selektovanje svih pojavljivanja jednog ili više čvorova iz banke drveta sa određenim osobinama
- Podrška za višeslojne i poravnate banke drveta
- Kvantifikovani i negirani podupiti (kao u "*nađi sve klauzule tačno sa tri objekta i bez subjekta*")
- Referisanje između čvorova ("*nađi roditeljski čvor i dete-čvor koji imaju isti padež i rod, ali različit broj*")
- Tekstualna i grafička reprezentacija upita odgovaraju strukturi rezultujućeg poddrveta
- Podjezik za generisanje izveštaja nad rezultujućim drvetima
- Podrška za regularne izraze, osnovne aritmetičke operacije i operacije nad niskama


LINDAT
Repository
TreeQuery
Treex
More Apps
Events
About
CLARIN
PML Tree Query
Select Treebank
Previous Queries
Documentation
Login
Available Treebanks
Filter treebanks ...
Show only accessible treebanks
Reset filter
Select Prague Dependency Treebank 3.0
Select The Penn Treebank 3, WSJ data set
Select Prague Czech-English Dependency Treebank - Czech Part Only
Select The Penn Treebank 3, Atis data set
Select The Penn Treebank 3, Brown data set, parsed and tagged
Select The Penn Treebank 3, Switchboard data set, parsed, tagged, and dysfluency annotated
Select Prague Dependency Treebank 2.5
Select Sample of The Prague Dependency Treebank 2.0
Select BNC Sample
Select Bulgarian Treebank

Primer 1:

U primeru je dat upit koji vraća frekvencije funkcija

```
nonterminal $n := []
        >> for $n.functions
                give $1, count()
                sort by $2 desc
```

("*Izaberi sve neterminalne čvorove. Uzmi njihove funkcije i prebroj koliko puta se svaka od njih pojavljuje, sortiraj ih po broju pojavljivanja*")

Kao izlaz dobija se sledeća lista:

738953
SBJ 116577
TMP 27189
LOC 19919
PRD 19793
CLR 18345

Primer 2:

- Prikazuje sve tipove polileksičkih jedinica

t-root $r := [ ];

>> for $r.mwes/type give

$1,count() sort by $2 desc

PML Tree Query

Select Treebank Previous Quer

Login

Relations Node Types Attributes Operators Functions

```
t-root $r := [ ];
        >> for $r.mwes/type give
        $1,count() sort by $2 desc
```

Query w/o Filters Visualize Clear

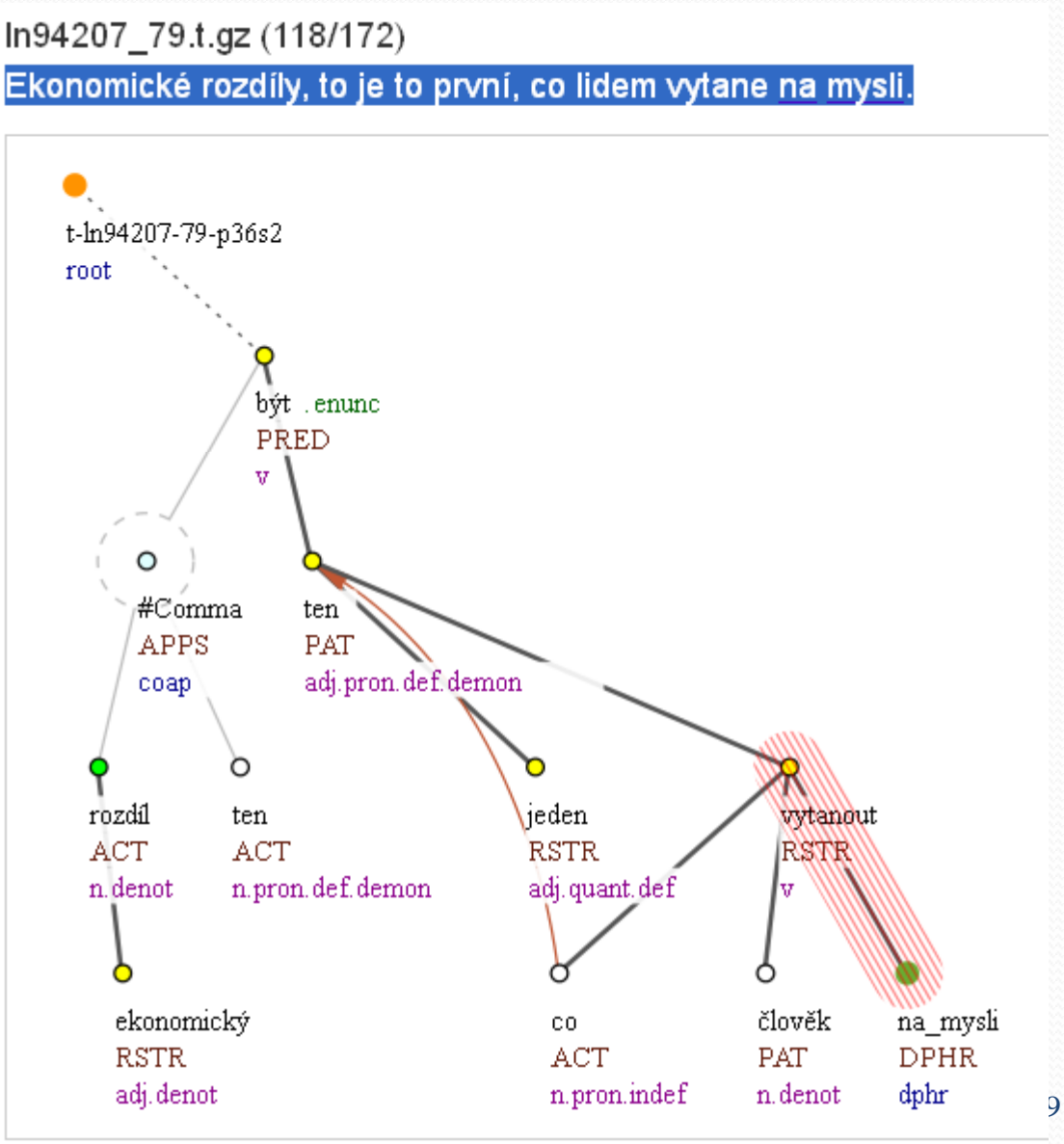
Result: 11 rows

$1	$2
	20495
lexeme	20078
person	6927
institution	4940
number	3629
object	2883
time	2644
location	1876
address	136
foreign	132
biblio	35

Ekonomske razlike, to je prvo što ljudima pada na pamet.

Primer 3:

```
t-node $n3 :=
[ functor = "DPHR”,
    same-tree-as t-root
    [ member mwes
    [ tnode.rfs $n3 ]
] ];
```



Hvala na pažnji!