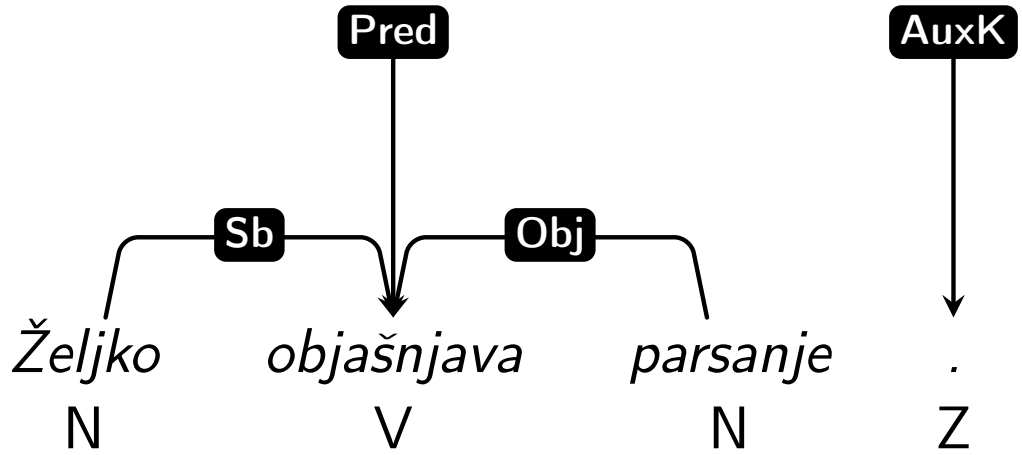
Pristupi ovisnosnom parsanju hrvatskih tekstova

Željko Agić

Sveučilište u Zagrebu
Filozofski fakultet
Odsjek za informacijske i komunikacijske znanosti

2012-07-09

Pregled



Pregled

Pitanja

- Što je ovisnosno parsanje teksta?
- Kako i zašto ovisnosno parsati tekst računalom?
- Kako što točnije i učinkovitije ovisnosno parsati tekstove pisane hrvatskim jezikom?

Hipoteze

- Tekstovi pisani hrvatskim jezikom mogu se robustno, točno i učinkovito ovisnosno parsati.
- Točnost ovisnosnoga parsanja može se povećati uporabom jezičnih resursa za hrvatski jezik, bez gubitka robusnosti i učinkovitosti.

Sadržaj

- ovisnosno parsanje
  - definicija parsanja
  - parser kao inteligentni računalni sustav
  - parsanje jezika i parsanje teksta
  - opći model parsera teksta
  - ovisnosno parsanje i ovisnosni parser
- postojeći pristupi
  - natjecanja u ovisnosnom parsanju
  - ovisnosni parseri temeljeni na teoriji grafova
  - prijelazнički ovisnosni parseri
- jedan model ovisnosnog parsera hrvatskih tekstova
  - neki pristupi poboljšavanju parsera
  - predloženi pristup
- zaključak
- nacrt budućih istraživanja

Ovisnosno parsanje

Sintaktička analiza — parsanje — neke definicije

- Parsanje je sintaktička analiza.
- Sintaktički analizirati znači provesti analizu s gledišta sintakse.
- Sintaksa je
  - jezikoslovna disciplina ili razina jezičnoga opisa,
  - skup pravila za opis nekoga jezika na toj razini jezičnoga opisa,
  - instancija tih pravila nad nekim jezičnim uzorkom, itd.
- Sintaktička analiza je analiza uloga riječi i skupova riječi u rečenicama nekog jezika prema nekom sintaktičkom formalizmu.
- Sintaktička analiza naziva se *parsanje* iz povijesnih razloga.

Ovisnosno parsanje

Elementi rečeničnoga ustroja. Jednostavne i složene rečenice

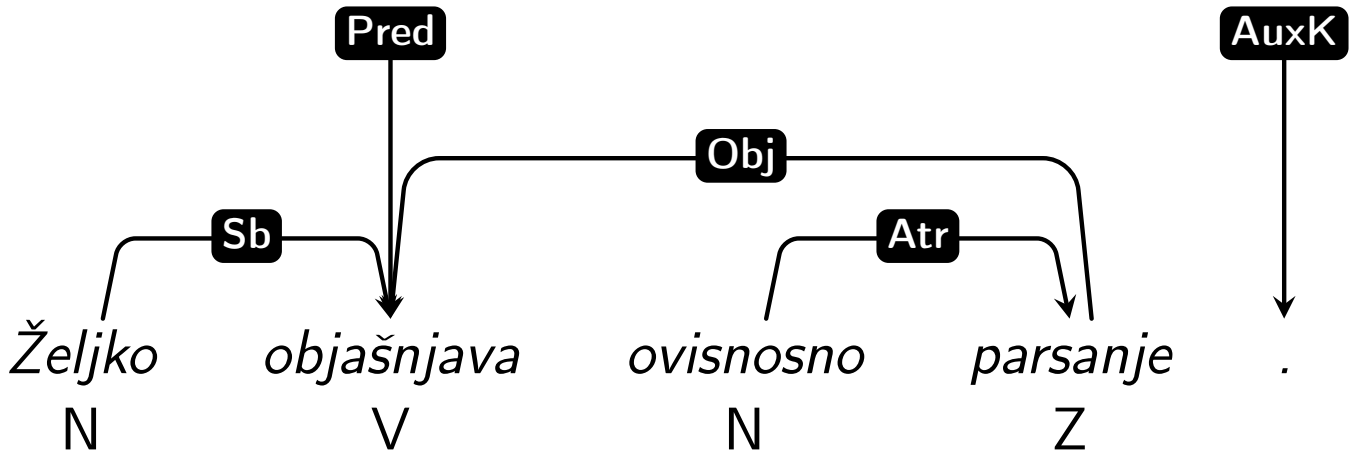
Sintaktička analiza je analiza ***uloga riječi i skupova riječi u rečenicama*** *nekog jezika prema nekom sintaktičkom formalizmu.*

- elementi rečeničnoga ustroja jednostavne rečenice
  - samostalni elementi
    - predikat — radnja, subjekt — vršitelj, objekt — trpitelj
    - priložna oznaka — dodatni opis radnje
      Željko je igrao nogomet svake srijede.
  - nesamostalni elementi
    - atribut i apozicija — dodatni opisi vršitelja i trpitelja
      Amater Željko je igrao loš nogomet svake srijede.
- složene rečenice
  - nezavisno-složene — koordinacija
    Željko je igrao nogomet, a vani je padala kiša.
  - zavisno-složene — subordinacija
    Željko je igrao nogomet dok je vani padala kiša.

Ovisnosno parsanje

Uvođenje elemenata rečeničnoga ustroja u rečenicu

Elementi rečeničnoga ustroja se uvode u rečenicu jedni po drugima.

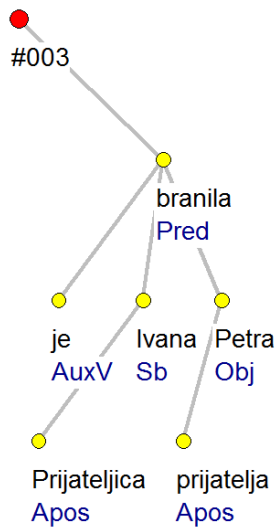


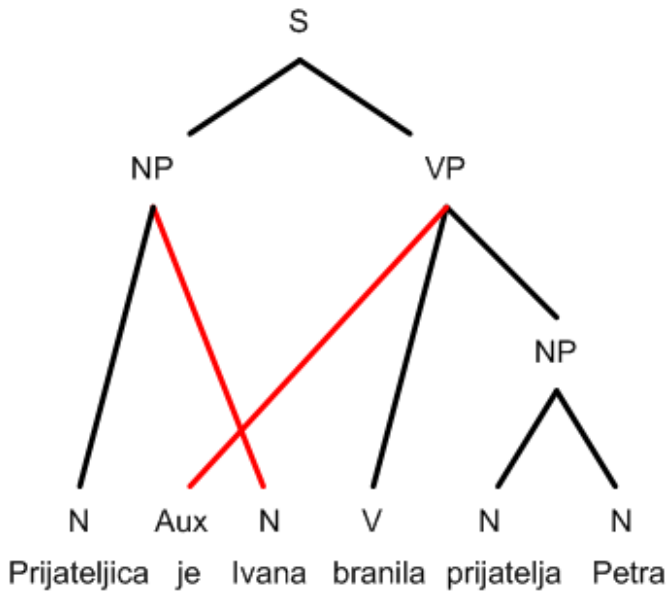
Sintaktički formalizam podrazumijeva opis elemenata rečeničnoga ustroja i opis načina njihovoga uvođenja u rečenicu.

Ovisnosno parsanje

Sintaktički formalizmi

- sintaktički formalizmi po opisu uvođenja
  - sintaksa fraznih struktura (en. *phrase structure*, *constitutency*)
  - ovisnosna sintaksa (en. *dependency*)



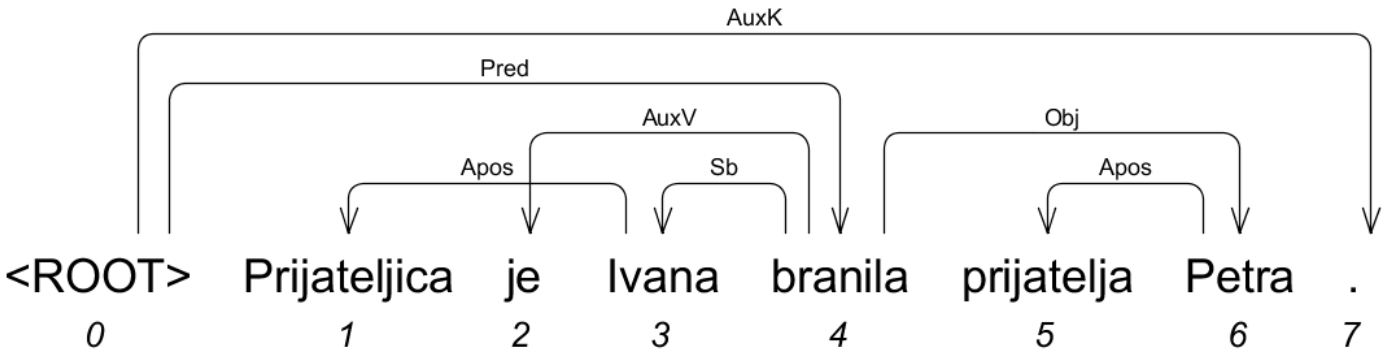


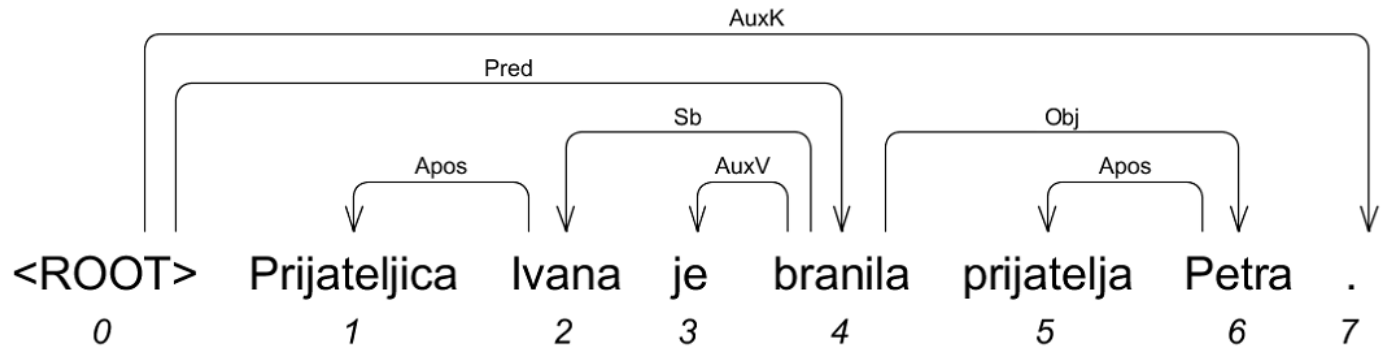
Ovisnosni modeli sintakse smatraju se prikladnijima za jezike sa slobodnijim redoslijedom riječi.

Ovisnosno parsanje

Sintaktički formalizmi — projektivnost i neprojektivnost

- (ne)projektivnost se odnosi na svojstvo pojedinih elemenata rečeničnoga ustroja da (ne) predstavljaju neprekinute slijedove riječi

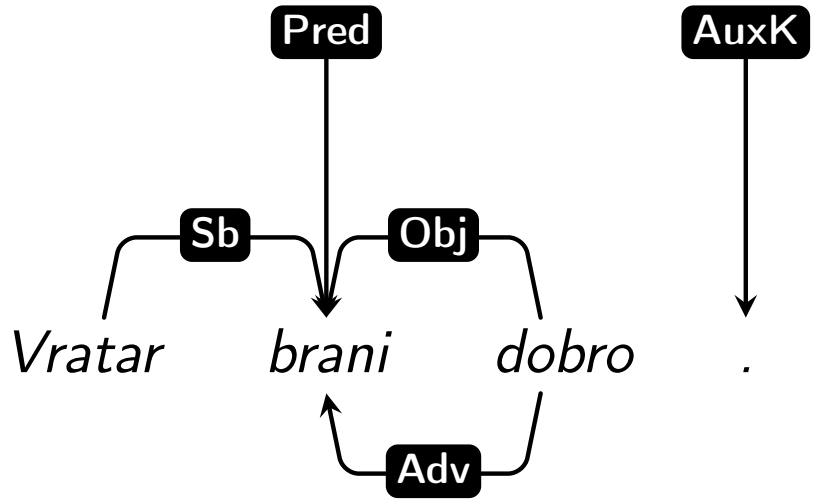




Ovisnosno parsanje

Sintaktička višeznačnost prirodnoga jezika

Jezik je višeznačan na svim razinama jezičnoga opisa.
Višeznačnost jezika postoji s razlogom — **olakšava** razmjenu obavijesti.



Koju obavijest ljudi usvajaju iz ove rečenice i kako to rade?
Kako u računalo ugraditi znanje za takvu vrstu obradbe?

Ovisnosno parsanje

Parser kao inteligentni računalni sustav

Parser je inteligentni računalni sustav kojim se provodi sintaktička analiza rečenica nekoga jezika u skladu sa zadanim sintaktičkim formalizmom.

- umjetna inteligencija
  - stvaranje strojeva koji usporedivo dobro izvršavaju zadatke za koje ljudi koriste inteligenciju kad ih izvršavaju
- parsanje jezika i parsanje teksta
  - generativni sintaktički model — formalne gramatike, formalni jezici
    - parseri formalnom gramatikom (CYK, Earley, itd.)
  - implicitni model — obradba prirodnoga jezika
    - parseri temeljeni na ručno izrađenim pravilima
    - parseri temeljeni na podatcima

Ovisnosno parsanje

Tražena svojstva parsera prirodnoga jezika

- robustno razrješavanje višeznačnosti

  Parser je robustan ako svakoj rečenici dodijeli barem jedno parsno stablo.

  Parser razrješuje sintaktičku višeznačnost ako svakoj rečenici dodijeli najviše jedno parsno stablo.

- točnost

  Parser je potpuno točan ako svakoj rečenici dodijeli baš ono parsno stablo koje predstavlja točno tumačenje te rečenice prema zadanome formalizmu.

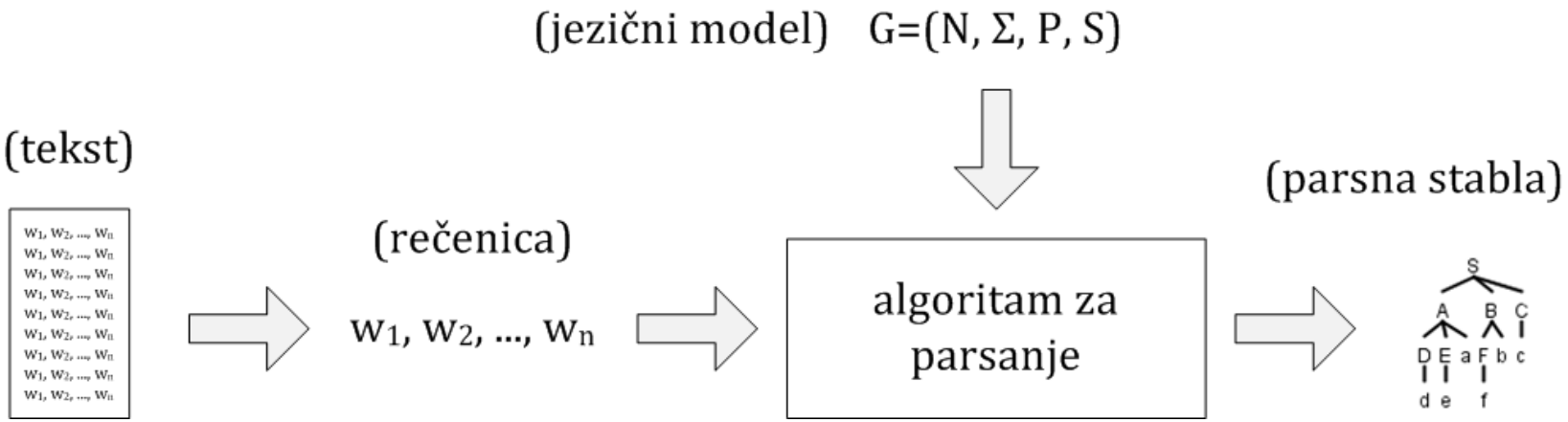
- učinkovitost

  Parser je potpuno učinkovit ako rečenice parsa u linearnom vremenu.

Ovisnosno parsanje

Tražena svojstva, parseri gramatikom i parseri teksta

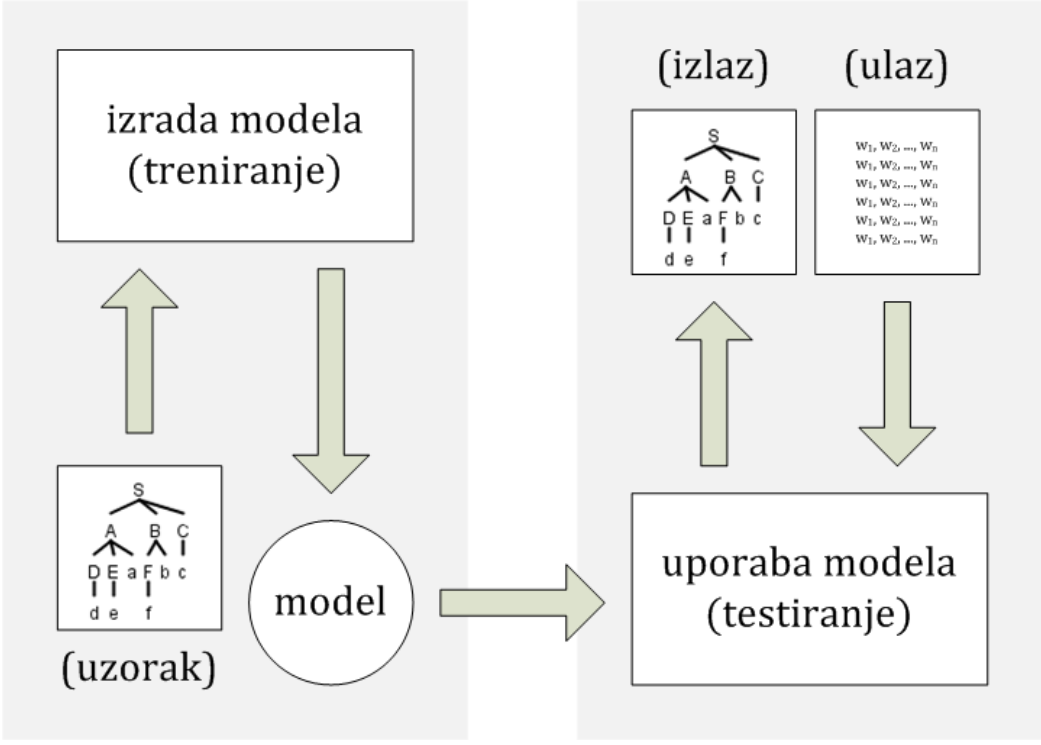
- parseri gramatikom ne zadovoljavaju neka od svojstava
  - nijedan ne razrješuje višeznačnost robustno
  - problem (ne)pokrivenosti prirodnoga jezika formalnom gramatikom
  - problem nemogućnosti razrješivanja postojećim parserima
- opći model parsera teksta
  - jezični model i algoritam za parsanje



Ovisnosno parsanje

Parseri teksta temeljeni na jezičnim resursima

- dva razdvojena pristupa
  - ovisnosni parseri temeljeni na ručno izrađenim pravilima
    - često se nazivaju i parserima temeljenima na gramatikama
    - zadržavaju probleme pokrivenosti i posljedične nemogućnosti robustnoga razrješivanja sintaktičke višeznačnosti
  - ovisnosni parseri temeljeni na podatcima



Ovisnosno parsanje

Sintaktički označeni računalni korpusi — banke ovisnosnih stabala

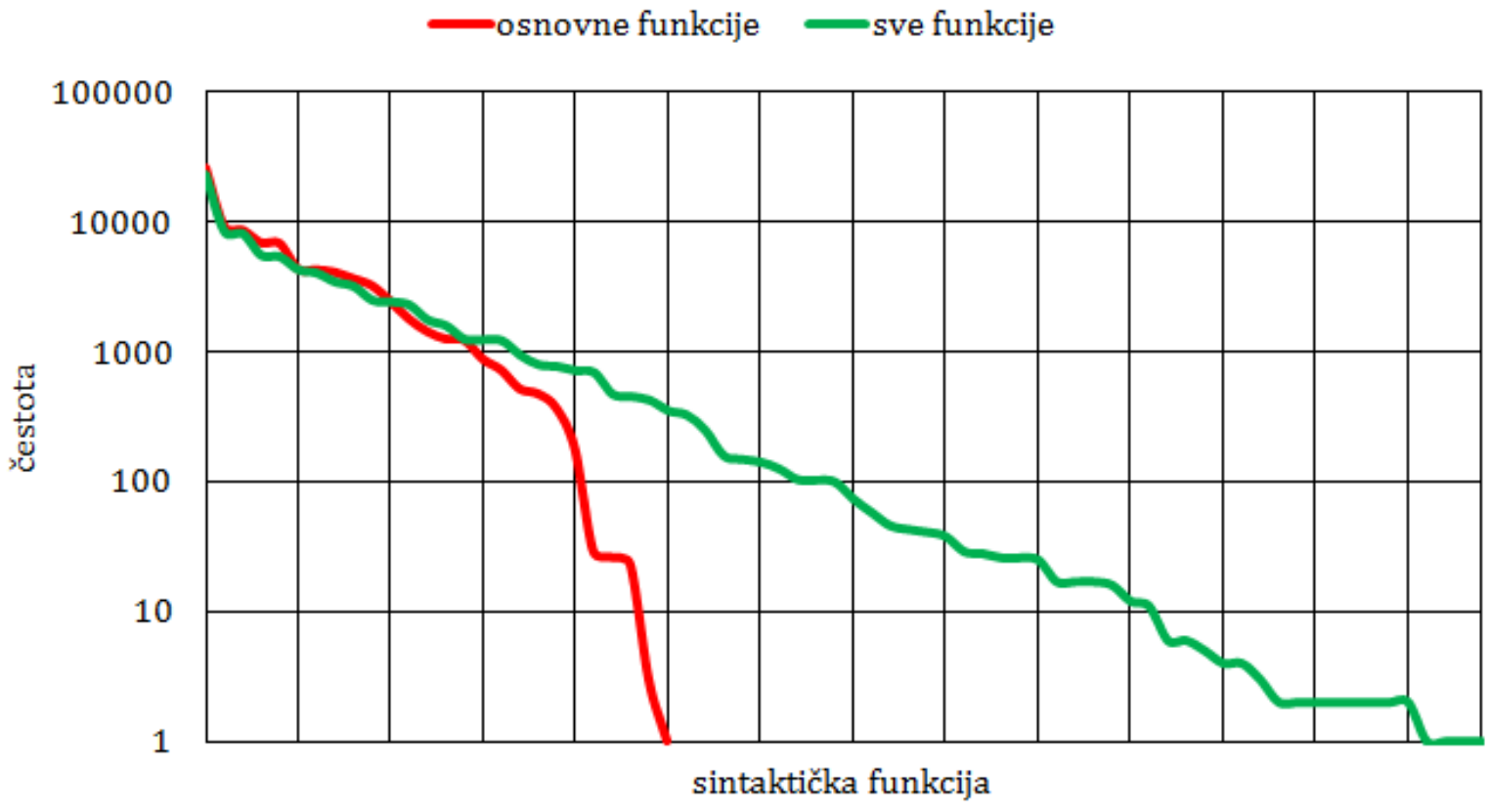
- banka ovisnosnih stabala
  - korpus tekstova pisanih nekim jezikom
  - označene granice rečenica i riječi
  - svakoj rečenici dodijeljeno ovisnosno stablo
  - najčešće također lematiziran i morfosintaktički označen
- Hrvatska ovisnosna banka stabala — HOBS
  - izgrađuje se nad novinskim korpusom CW100 (cca 108 kw)
  - slijedi načela izgradnje Praške banke ovisnosnih stabala

značajka	broj
rečenica	3,465
pojavnica	88,045
oblik	20,703
lema	10,481 (10,527)
morfosintaktička oznaka	828
sintaktička funkcija	26 (69)

Ovisnosno parsanje

Hrvatska ovisnosna banka stabala

- čestota sintaktičkih funkcija
  - osnovne i proširene, odnosno sve funkcije



Ovisnosno parsanje

Hrvatska ovisnosna banka stabala

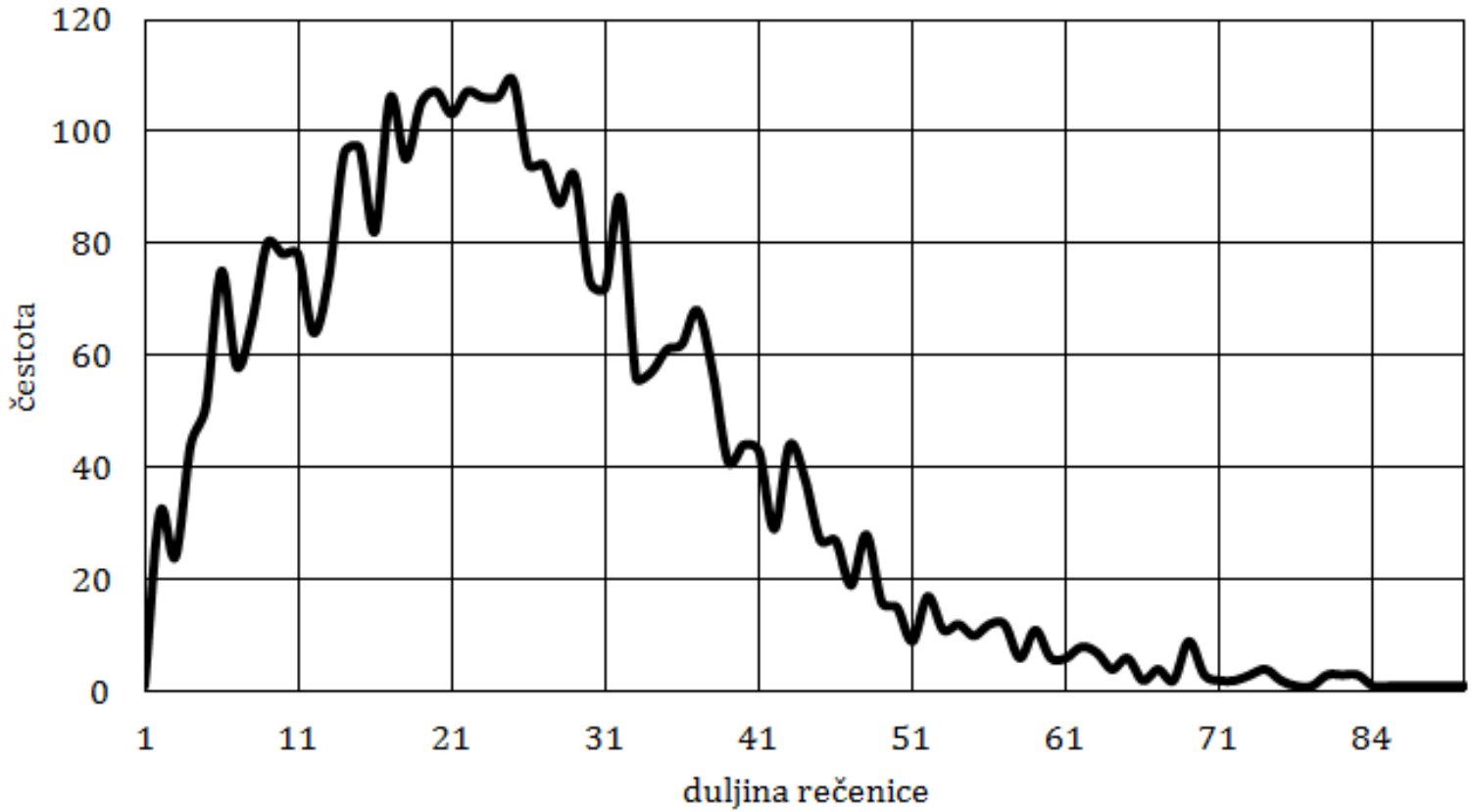
- čestota sintaktičkih funkcija
  - izdvojene samo osnovne sintaktičke funkcije

Atr	**Adv**	**AuxP**	**Sb**	**Obj**	**AuxX**	**Pred**
25816	9150	8607	6874	6776	4316	4255
AuxV	**Coord**	**AuxK**	**AuxG**	**AuxC**	**Pnom**	**AuxZ**
4075	3627	3219	2441	1782	1422	1248
ExD	**AuxY**	**AuxR**	**Apos**	**AuxT**	**Atv**	**AtvV**
1225	876	721	516	473	366	178
AtrAdv	**AuxO**	**AdvAtr**	**AtrObj**	**ObjAtr**		
29	26	23	3	1		

Ovisnosno parsanje

Hrvatska ovisnosna banka stabala

- čestota rečenica po broju pojavnica
  - najviše rečenica između 15 i 25 pojavnica



Ovisnosno parsanje

Hrvatska ovisnosna banka stabala

- razdioba sintaktičkih funkcija po vrstama riječi
  - izdvojene osnovne sintaktičke funkcije s ozbirom na definiciju elemenata rečeničnoga ustroja i najčešće vrste riječi

	pridjev (A)	veznik (C)	broj (M)	imenica (N)	zamjenica (P)	prilog (R)	prijedlog (S)	glagol (V)
Adv	299	127	359	4800	401	2421	85	647
Apos	1	17	0	1	4	33	1	3
Atr	9477	4	752	11209	1586	221	5	1644
AuxC	0	1517	0	2	126	57	28	5
AuxP	2	100	0	190	0	46	8260	1
Coord	0	3141	0	2	14	32	6	1
Obj	120	2	138	3644	860	42	1	1927
Pnom	517	0	37	670	32	59	2	102
Pred	63	0	0	1	3	0	0	4188
Sb	82	1	196	4853	1179	41	2	337

Ovisnosno parsanje

Hrvatska ovisnosna banka stabala

- HOBS nije dovršen resurs
  - prilagodbe preuzetog formalizma posebnostima hrvatskih tekstova
  - ispravljanje pogrešaka
  - sustavno označavanje složenih rečenica
  - označavanje čitavog korpusa CW100 — preostalo je 1,161 rečenica
  - uvođenje novih tekstova
    - nastavak ručnog označavanja
    - hr-si paralelni korpus *1984.* iz projekta MTE
    - poluautomatsko prebacivanje ovisnosnih stabala
- eksperimenti s ovisnosnim parsanjem drugih jezika
  - korištene banke stabala od min. 30 kw do max. 0.5 mw
  - HOBS dovoljno velik za treniranje ovisnosnih parsera i preliminarno testiranje

Postojeći parseri

Odabir pristupa parsanju za testiranje na HOBS-u

- testirati ovisnosne parsere na hrvatskim tekstovima iz HOBS-a
- brojni javno dostupni ovisnosni parseri
  - MaltParser, MSTParser, ISBN Parser, DeSR, kMST Parser, itd.
  - velik interes za ovisnosno parsanje u zadnjih 10-ak godina
- pristup odabiru
  - isprobati različite paradigme
  - isključiti neisplative parsere iz testiranja
- natjecanja u ovisnosnom parsanju na skupu CoNLL 2006. i 2007.
  - izdvaja se MaltParser i MSTParser
  - parseri temeljeni na podatcima i teoriji grafova, odnosno prijelazničkim sustavima

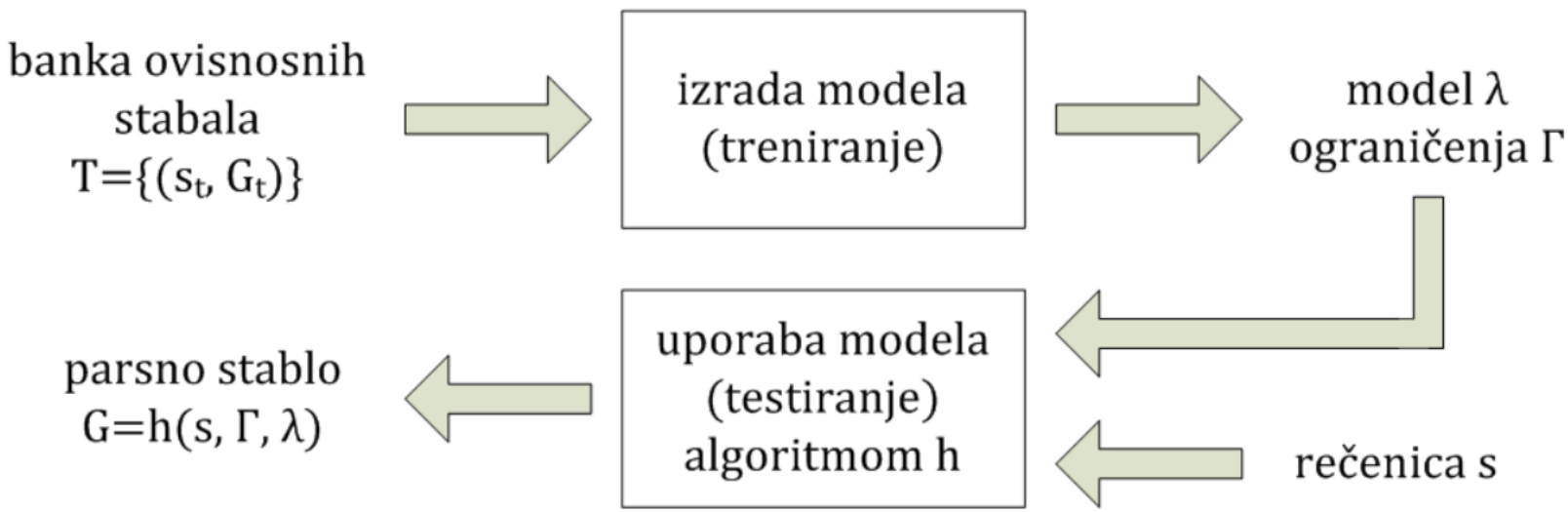
	ara	kin	češ	dan	niz	njem	jap	por	slo	špa	šve	tur	ukupno
MST	**66.9**	85.9	**80.2**	**84.8**	**79.2**	**87.3**	**90.7**	**86.8**	**73.4**	**82.3**	82.6	63.2	**80.3**
Malt	**66.7**	86.9	78.4	**84.8**	**78.6**	85.8	**91.7**	**87.6**	70.3	**81.3**	**84.6**	**65.7**	**80.2**

Postojeći parseri

Osnovne postavke problema

- ovisnosno parsanje je optimizacijski problem
- opći model ovisnosnog parsera temeljenoga na podatcima
- jezični model, parsni algoritam, treniranje i testiranje

$$M = (\Gamma, \lambda, h), G = h(s, \Gamma, \lambda)$$



Postojeći parseri

Ovisnosno parsanje temeljeno na teoriji grafova

- ovisnosno stablo je graf
  - svojstvo posjedovanja korijenskog čvora, povezanosti, usmjerenosti, acikličnosti, jedne glave po relaciji
- primjena metoda iz teorije grafova
  - jezični model sadrži preferencije povezivanja pojedinih riječi u relacije uz dodjelu pojedinih sintaktičkih funkcija
  - preferencije su definirane jezičnim značajkama
  - koriste se algoritmi za pronalaženje najvećeg prostirućeg (razapinjućeg) stabla (en. *maximum spanning tree*, MST)
- neusmjereni algoritmi, globalno pretraživanje, ograničene globalne značajke (en. *arc-factored*)
- predstavnik — generator parsera MSTParser
  - jezični modeli prvog i drugog reda
  - algoritmi za projektivno (Eisner) i neprojektivno (Chu-Liu-Edmonds) parsanje
  - parsanje u kvadratnom i kubnom vremenu

Postojeći parseri

Ovisnosno parsanje temeljeno na prijelazničkim sustavima

- prijelazнички sustav je formalni automat
  - određen s pomoću skupa konfiguracija (ili stanja) i funkcije koja, najčešće ovisno o nekome ulazu, određuje njegovo prelaženje iz jedne u drugu konfiguraciju (ili iz jednoga stanja u drugo)
  - u ovisnosnom se parsanju najčešće koristi stog i ulazna vrpca — potisni automat
- parsanje u linearnom vremenu
  - algoritam po zadanoj strategiji pita jezični model o idućem prijelazu
  - izrada jezičnoga modela i odabir značajki najvažniji su koraci u uporabi prijelazničkih parsera
- usmjereni algoritmi, lokalno pretraživanje, lokalne značajke
- predstavnik — generator parsera MaltParser
  - pet razreda algoritama, devet različitih algoritama
  - veliki broj postavki
  - sustav MaltOptimizer za odabir postavki prema značajkama banke ovisnosnih stabala

Postojeći parseri

Mjere za vrjednovanje ovisnosnih parsera

- formalni kriteriji za vrjednovanje
  - preduvjeti — robustno razrješavanje višeznačnosti
  - optimizacijski kriteriji — točnost i učinkovitost
- mjere za vrjednovanje točnosti
  - povezivanje pojavnica uz dodjelu sintaktičkih funkcija (en. *labeled attachment score*, LAS)
  - povezivanje pojavnica bez dodjele sintaktičkih funkacija (en. *unlabeled attachment score*, UAS)
  - dodjela sintaktičkih funkcija pojavnicama (en. *label attachment*, LA)
  - preciznost i odziv pri dodjeli pojedinih sintaktičkih funkcija
  - pojedine mjere s obzirom na vrstu riječi, morfosintaktičke značajke, svojstva ovisnosnih stabala, itd.
- mjere za vrjednovanje učinkovitosti
  - vrijeme izvođenja i memorijski zahtjevi postupaka treniranja i testiranja parsera

Postojeći parseri

Postavke eksperimenta

- usklađenost s natjecanjima CoNLL 2006. i 2007.
- skup za testiranje modela od cca 5,000 pojavnica
- deseterostruka unakrsna provjera (en. *tenfold cross-validation*)
- korištene osnovne sintaktičke funkcije
- vrjednovano ukupno 11 ovisnosnih parsera iz generatora parsera MaltParser i MSTParser

značajka	**skup za treniranje**		**skup za testiranje**	
rečenica	3261.18 ± 4.20		203.82 ± 4.20	
pojavnica	82865.88 ± 6.87		5179.12 ± 6.87	
oblik	19927.06 ± 15.71		2594.06 ± 12.26	
lema	10166.00 ± 9.19		1909.00 ± 14.12	
morfosintaktička oznaka	817.94 ± 1.40		368.35 ± 4.41	
sintaktička funkcija	69.00 ± 0.00	26.00 ± 0.00	48.12 ± 0.84	23.24 ± 0.43

Postojeći parseri

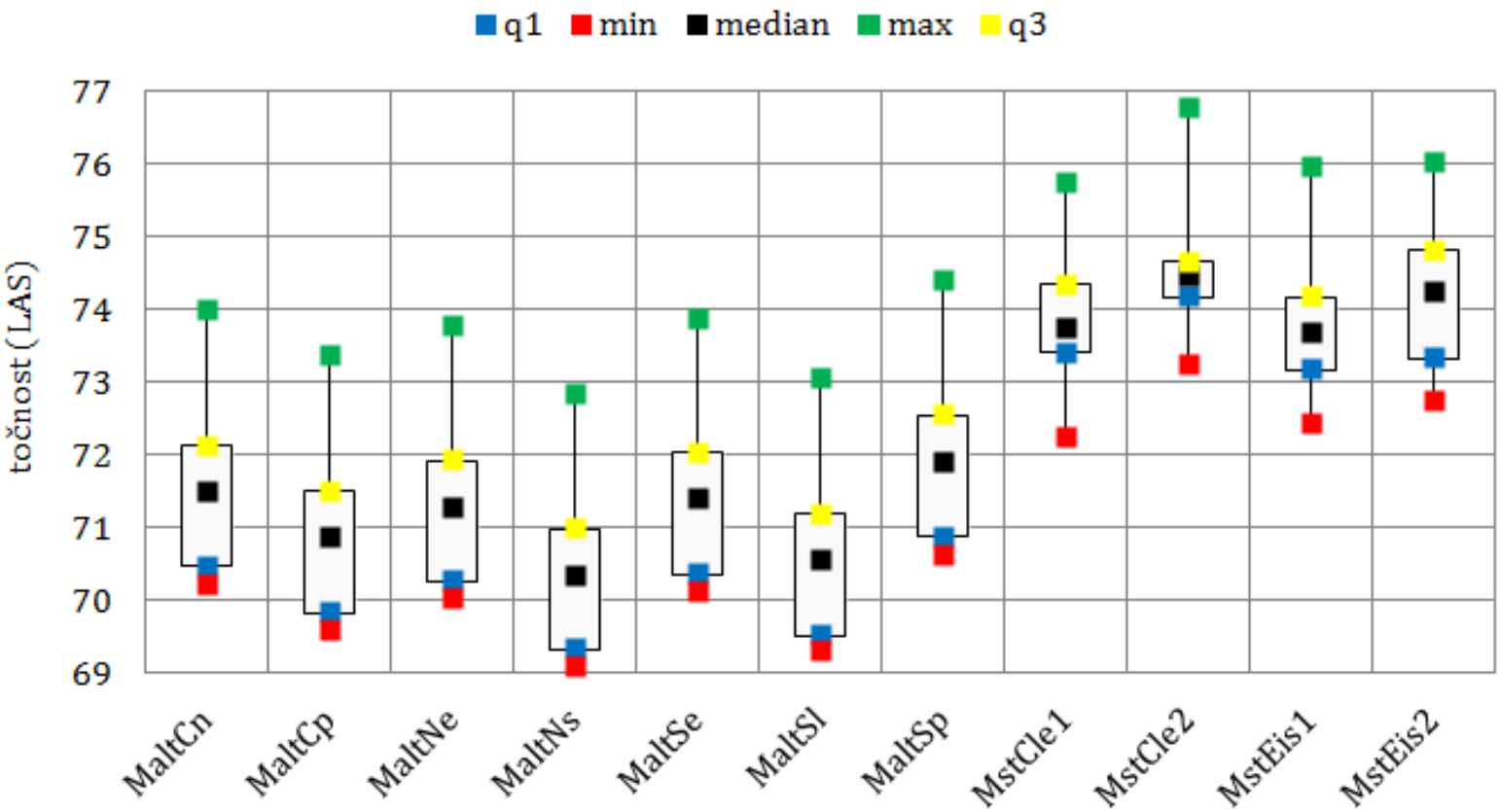
Rezultati eksperimenta — točnost parsera prema mjerama LAS, UAS i LA

parser	LA	LAS	UAS
MaltNe	83.74 ± 0.46	71.29 ± 0.74	77.13 ± 0.71
MaltNs	83.16 ± 0.47	70.35 ± 0.73	76.44 ± 0.70
MaltCp	83.46 ± 0.48	70.87 ± 0.73	76.80 ± 0.69
MaltSp	**84.05 ± 0.44**	**71.91 ± 0.74**	**77.59 ± 0.73**
MaltCn	**83.88 ± 0.46**	**71.50 ± 0.74**	**77.30 ± 0.72**
MaltSe	83.75 ± 0.42	71.39 ± 0.73	77.23 ± 0.72
MaltSl	83.28 ± 0.48	70.56 ± 0.73	76.54 ± 0.71
MstEis1	85.57 ± 0.36	73.73 ± 0.65	80.92 ± 0.61
MstEis2	**85.64 ± 0.39**	**74.17 ± 0.64**	**81.27 ± 0.59**
MstCle1	85.76 ± 0.35	73.88 ± 0.58	80.99 ± 0.50
MstCle2	**85.87 ± 0.38**	**74.53 ± 0.57**	**81.69 ± 0.44**

Postojeći parseri

Rezultati eksperimenta — točnost parsera prema mjerama LAS, UAS i LA

- parseri temeljeni na teoriji grafova bolji od prijelazničkih parsera
- najbolji sustav s CLE MST algoritmom, cca 74.53 LAS



Postojeći parseri

Rezultati eksperimenta — točnost parsera s obzirom na vrstu riječi

parser	mjera	A	C	M	N	P	R	S	V	Z
MaltCn	LA	91.32	77.88	70.21	82.08	79.71	79.73	95.43	77.55	88.66
	LAS	88.00	51.92	61.09	73.94	75.00	65.70	69.59	65.18	71.40
	UAS	89.88	56.37	73.60	83.29	85.75	73.08	70.41	70.50	73.83
MaltSp	LA	91.38	77.36	**70.68**	82.36	79.48	79.90	95.62	78.05	88.78
	LAS	88.02	51.99	**61.23**	74.45	75.12	65.87	69.79	66.13	71.79
	UAS	89.82	56.32	73.43	83.73	86.18	72.67	70.58	71.02	74.21
MstCle2	LA	**92.96**	87.94	68.19	81.84	**80.79**	**80.77**	98.57	**80.65**	**91.13**
	LAS	**89.96**	62.09	59.99	**74.50**	**76.08**	**68.19**	**74.72**	**71.81**	73.26
	UAS	**92.69**	64.31	76.39	**86.60**	**89.22**	**77.84**	75.35	**79.11**	75.40
MstEis2	LA	92.25	**88.12**	67.69	**81.85**	79.92	80.45	**98.58**	80.54	90.78
	LAS	88.73	**62.12**	61.00	74.33	74.32	66.81	74.63	71.54	**73.55**
	UAS	91.28	**64.34**	**77.59**	86.31	87.45	76.03	**75.36**	78.87	**75.61**

Postojeći parseri

Rezultati eksperimenta — točnost parsera s obzirom na sintaktičku funkciju

parser	mjera	Adv	Apos	Atr	AuxC	AuxP	Coord	Obj	Pnom	Pred	Sb
MaltCn	LAS	70.67	**45.88**	83.77	74.36	71.99	46.28	67.40	**66.55**	36.45	69.14
	UAS	83.16	**50.45**	88.40	75.81	72.46	46.92	79.94	70.33	43.82	76.73
MaltSp	LAS	**71.31**	44.73	**83.98**	**75.68**	72.08	46.96	68.15	66.35	37.33	70.12
	UAS	83.41	48.50	**88.59**	**77.10**	72.53	47.79	80.08	70.43	44.33	77.53
MstCle2	LAS	69.01	37.40	81.80	71.94	**74.35**	**56.49**	**69.38**	65.18	68.10	72.51
	UAS	**85.58**	43.48	87.78	74.07	**75.06**	**57.73**	**84.64**	**77.52**	75.36	**81.67**
MstEis2	LAS	68.38	39.34	81.46	73.21	74.15	55.05	68.29	62.47	**69.09**	**72.63**
	UAS	84.67	44.23	87.44	74.86	74.90	56.41	83.95	74.38	**76.06**	81.34

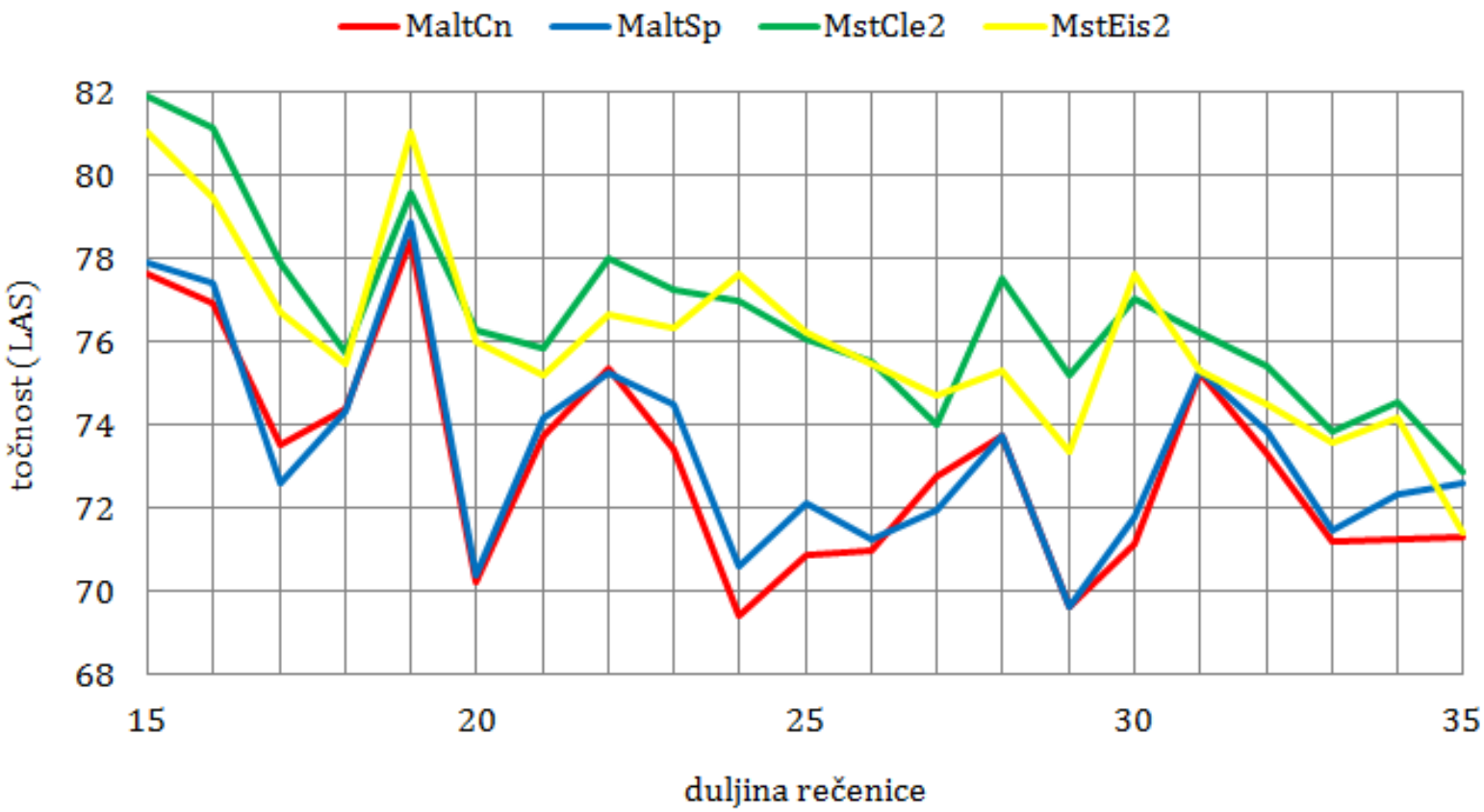
Postojeći parseri

Rezultati eksperimenta — preciznost i odziv dodjele sintaktičkih funkcija po LA

parser	mjera	Adv	Apos	Atr	AuxC	AuxP	Coord	Obj	Pnom	Pred	Sb
MaltCn	P	78.72	58.94	90.17	**94.04**	98.69	88.01	75.63	**69.95**	50.68	81.76
	O	77.24	36.63	91.48	83.90	94.50	67.14	72.55	49.35	82.69	83.35
MaltSp	P	**79.09**	58.70	**90.20**	93.67	**98.79**	87.96	**76.31**	69.46	50.89	82.32
	O	**77.50**	37.78	**91.76**	84.15	94.67	67.28	**73.57**	47.06	82.76	83.47
MstCle2	P	75.63	**64.29**	88.20	91.51	98.09	**90.22**	76.24	69.18	79.66	82.98
	O	76.04	**51.80**	91.69	88.81	**97.83**	83.74	71.37	54.36	**84.89**	**86.15**
MstEis2	P	75.44	60.07	87.97	92.31	97.96	89.37	75.71	66.29	**80.26**	**83.53**
	O	75.62	47.35	91.58	**89.44**	97.77	**84.87**	71.03	**57.10**	83.62	85.83

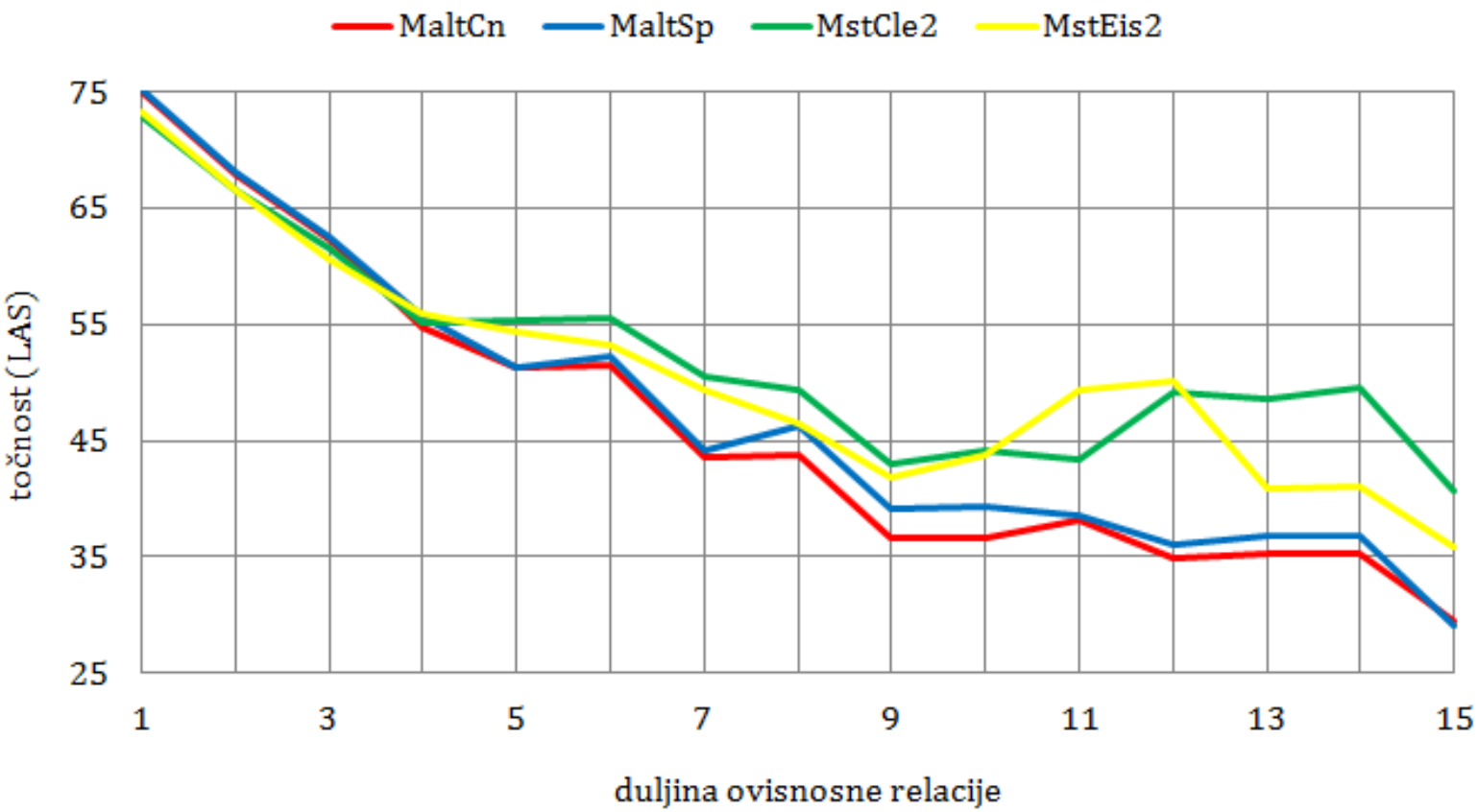
Postojeći parseri

Rezultati eksperimenta — točnost (LAS) s obzirom na duljinu rečenice



Postojeći parseri

Rezultati eksperimenta — točnost (LAS) s obzirom na udaljenost među pojavnicama



Postojeći parseri

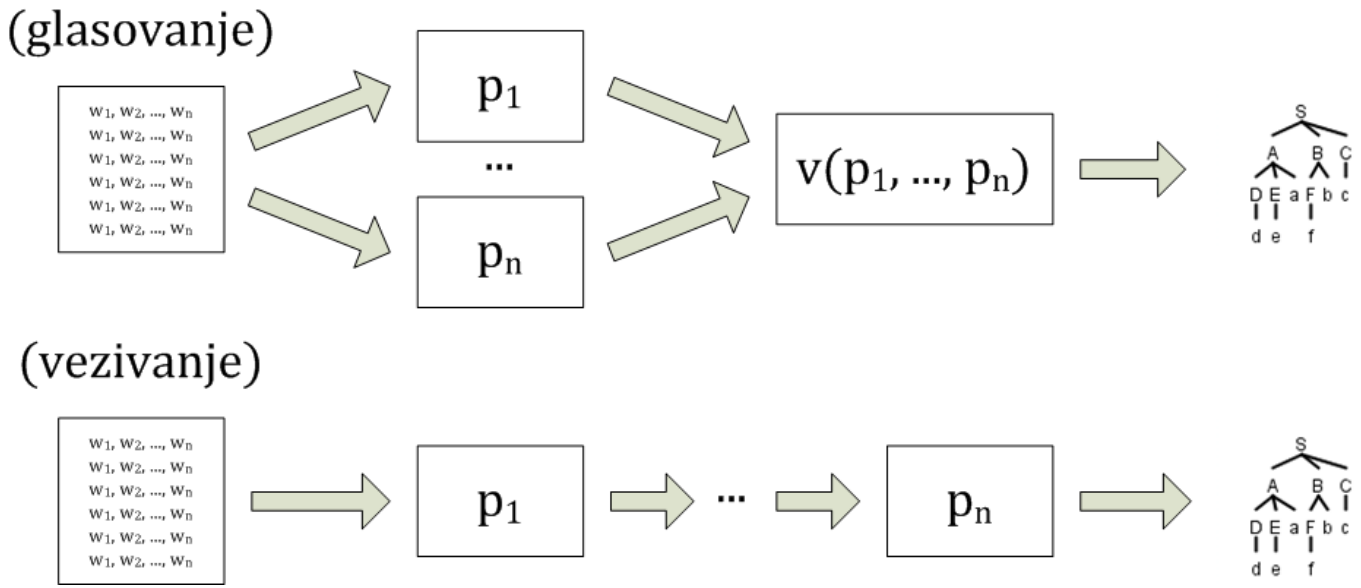
Rezultati eksperimenta — neki zaključci

- parseri temeljeni na podatcima primjenjivi za ovisnosno parsanje hrvatskih tekstova
- pristupi temeljeni na grafovima bolji od prijelazničkih pristupa
- točnost prema odabranim mjerama usporediva s točnošću istih parsera na natjecanjima CoNLL 2006. i 2007. na srodnim jezicima
  - češki 80.2 LAS, slovenski 73.4 UAS
  - HOBS cca 90 kw, SDT iz 2006. cca 30 kw — razlika u točnosti od cca 1.13 prema mjeri LAS?
- točnost povezivanja i označavanja obavijesno najvažnijih kategorija
  - predikat — 69.09 LAS, 76.06 UAS
  - subjekt — 72.63 LAS, 81.67 UAS
  - objekt — 69.38 LAS, 84.64 UAS
  - Kako povećati točnost povezivanja i označavanja ovih elemenata rečeničnoga ustroja u okviru ovisnosnoga parsanja temeljenoga na podatcima, odnosno na teoriji grafova?

Predloženi model

Pristupi povećavanju točnosti ovisnosnih parsera

- slaganje ovisnosnih parsera
  - glasovanje (en. *voting*) — paralelno
  - vezivanje (en. *stacking*) — serijski



- hibridizacija ovisnosnih parsera
  - uvođenje jezično-specifičnih modula
  - korištenje specifičnih jezičnih resursa
  - pitanje smislenosti pojedinih izbora s obzirom na prirodu problema

Predloženi model

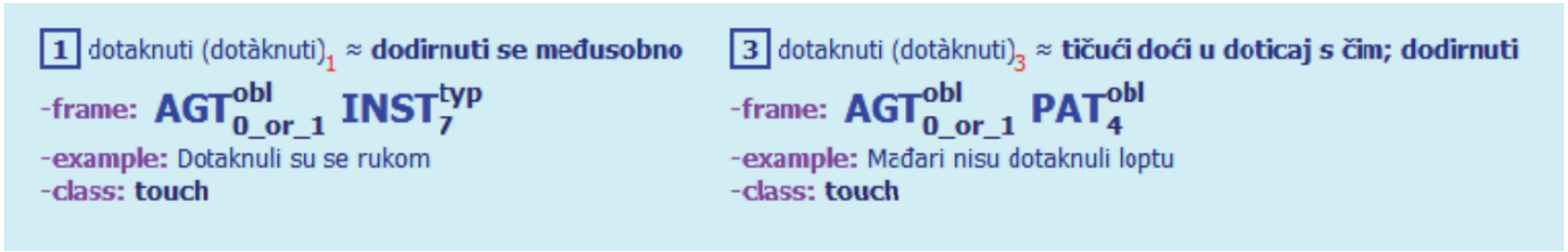
Pristupi povećavanju točnosti ovisnosnih parsera

- slaganje ovisnosnih parsera načelno daje mjerljiva poboljšanja
  - ovisno o polazišnoj točnosti pojedinih parsera
  - povezivanje raznorodnih parsera daje osjetnija poboljšanja
  - razlika među parserima temeljenima na grafovima i prijelazničkim parserima u prethodno prikazanome eksperimentu ne jamči značajnije poboljšanje rezultata
  - odnosi se na glasovanje i na vezivanje
- odabran hibridizacijski pristup
  - razvoj dodatnih modula temeljenih na pravilima je dugotrajan i narušava učinkovitost
  - ugradnja modula temeljenih na pravilima u postojeće paradigme je najčešće netrivijalna
  - korištenje dostupnih jezičnih resurasa za hrvatski jezik
    - valencijski rječnik glagola hrvatskoga jezika — CROVALLEX

Predloženi model

Valencijski rječnik hrvatskih glagola CROVALLEX

- valencija glagola (i drugih vrsta riječi) predstavlja model uvođenja elemenata u rečenicu preko ranije uvedenih elemenata i u osnovi je ovisnosnih teorija sintakse
- korištena inačica CROVALLEX-a 2.008
  - 1,797 lema glagola
  - 5,188 pripadajućih valencijskih okvira
  - svaki okvir uključuje podatak o broju mjesta koja se otvaraju za nove elemente rečeničnoga ustroja i traženim morfosintaktičkim svojstvima tih elemenata



Predloženi model

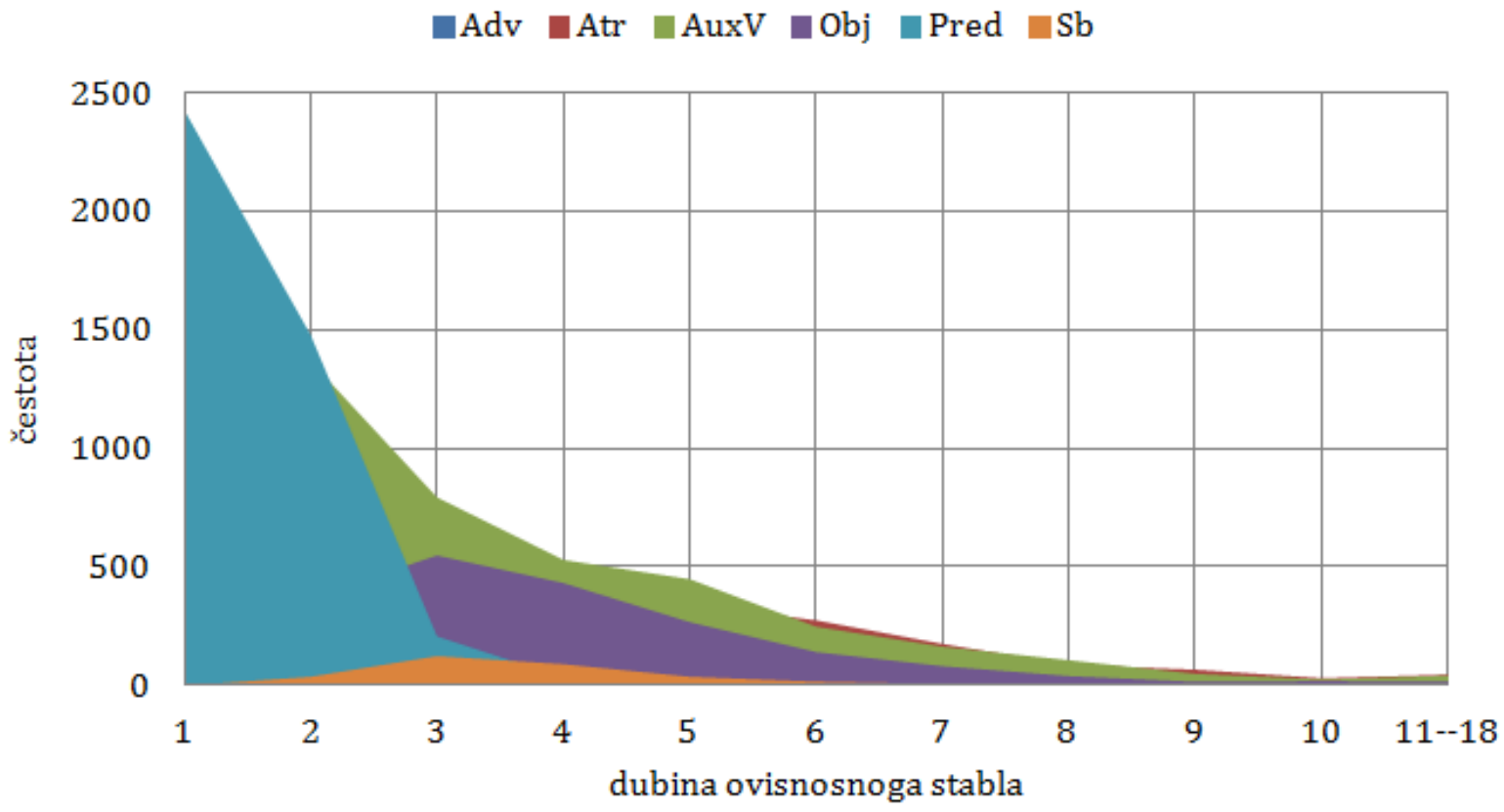
CROVALLEX i HOBS — pokrivenost glagola

- statička i dinamička pokrivenost
  - 1,525 lema i 12,958 pojavnih oblika glagola u HOBS-u (cca 15% od ukupnoga broja)
  - u CROVALLEX-u se nalazi cca 51.87% lema glagola iz HOBS-a
  - cca 45.64% lema glagola iz CROVALLEX-a nije se pojavilo u HOBS-u
  - CROVALLEX-om pokriveno cca 90.76% pojavnih oblika glagola iz HOBS-a (nepokrivenost cca 9.24%)
- visoka pokrivenost opravdava uporabu CROVALLEX-a u ovisnosnom parsanju
  - Kako ugraditi znanje o glagolima sadržano u CROVALLEX-u u postupak ovisnosnoga parsanja temeljenoga na grafovima?

Predloženi model

Uporaba CROVALLEX-a u ovisnosnom parsanju

- čestota sintaktičkih funkcija osnovnih elemenata rečeničnoga ustroja s obzirom na položaj u ovisnosnom stablu



Predloženi model

Uporaba CROVALLEX-a u ovisnosnom parsanju

- razdioba sintaktičkih funkcija pojavnica direktno ovisnih o predikatima

Sb	**AuxP**	**AuxV**	**Obj**	**Adv**	**AuxC**	**Pnom**
19.87%	16.38%	15.47%	12.17%	10.00%	5.34%	4.27%
Coord	**AuxR**	**AuxY**	**AuxX**	**AuxT**	**AuxG**	**Apos**
3.93%	2.01%	2.00%	2.00%	1.61%	1.42%	1.19%
AtvV	**Pred**	**ExD**	**AuxZ**	**AuxK**	**AuxO**	**Atr**
0.82%	0.65%	0.40%	0.35%	0.05%	0.05%	0.03%

Predloženi model

Hibridni ovisnosni parser temeljen na grafovima

- vrjednovanje predloženih ovisnosnih stabala valencijskim rječnikom
  - neka postoji neki broj kandidata za ovisnosno stablo neke rečenice hrvatskoga jezika
  - svaka ovisnosna relacija kojom se neka pojavnica vezuje uz glagolski predikat podložna je vrjednovanju CROVALLEX-om
  - trostupanjsko vrjednovanje ovisnih pojavnica
    - broj pojavnica
    - vrste riječi
    - morfosintaktička svojstva
  - dvostupanjsko rangiranje ovisnosnih stabala
    - prema statističkoj pouzdanosti (en. *k-best parsing*)
    - prema ocjeni iz CROVALLEX-a
- razviti novi ovisnosni parser temeljen na grafovima
  - daje k ovisnosnih stabala za svaku ulaznu rečenicu i svakom stablu pridružuje mjeru pouzdanosti
  - naknadno pridružuje mjere pouzdanosti s obzirom na CROVALLEX
  - stabla se nanovno rangiraju vrjednovanjem kombinacije dviju mjera

Predloženi model

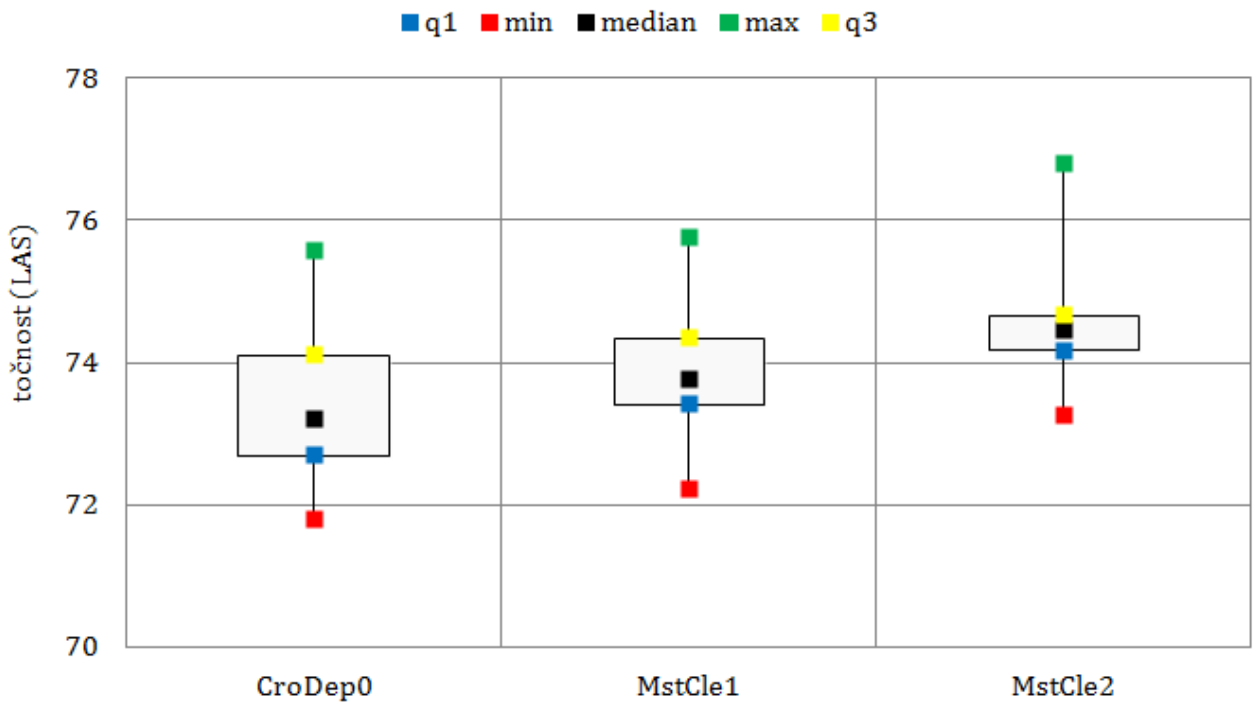
Hibridni ovisnosni parser temeljen na grafovima

- dvije razvojne faze
  - ovisnosni parser temeljen na grafovima
    - prototipni sustav radnoga naziva CroDep0
    - po uzoru na MSTParser
    - jezični model prvoga reda (en. *arc-factored*) i algoritam CLE
    - razvijen u programskom jeziku Java
  - *k-best* ovisnosno parsanje i uporaba CROVALLEX-a
    - prototipni sustav radnoga naziva CroDep
    - algoritam CLE neučinkovit za *k-best* parsanje
    - uporabljen algoritam kMST iz teorije grafova, provjeren u ovisnosnom parsanju engleskih tekstova
    - $k = 10$ u prototipnoj izvedbi
    - postojeća izvedba vrjednovana samo po mjeri UAS
    - dodana interakcija s jezičnim modelom za dodjelu sintaktičkih funkcija
    - razvijen modul za vrjednovanje ovisnosnih relacija CROVALLEX-om
    - također razvijen u programskom jeziku Java

Predloženi model

Postavke eksperimenta

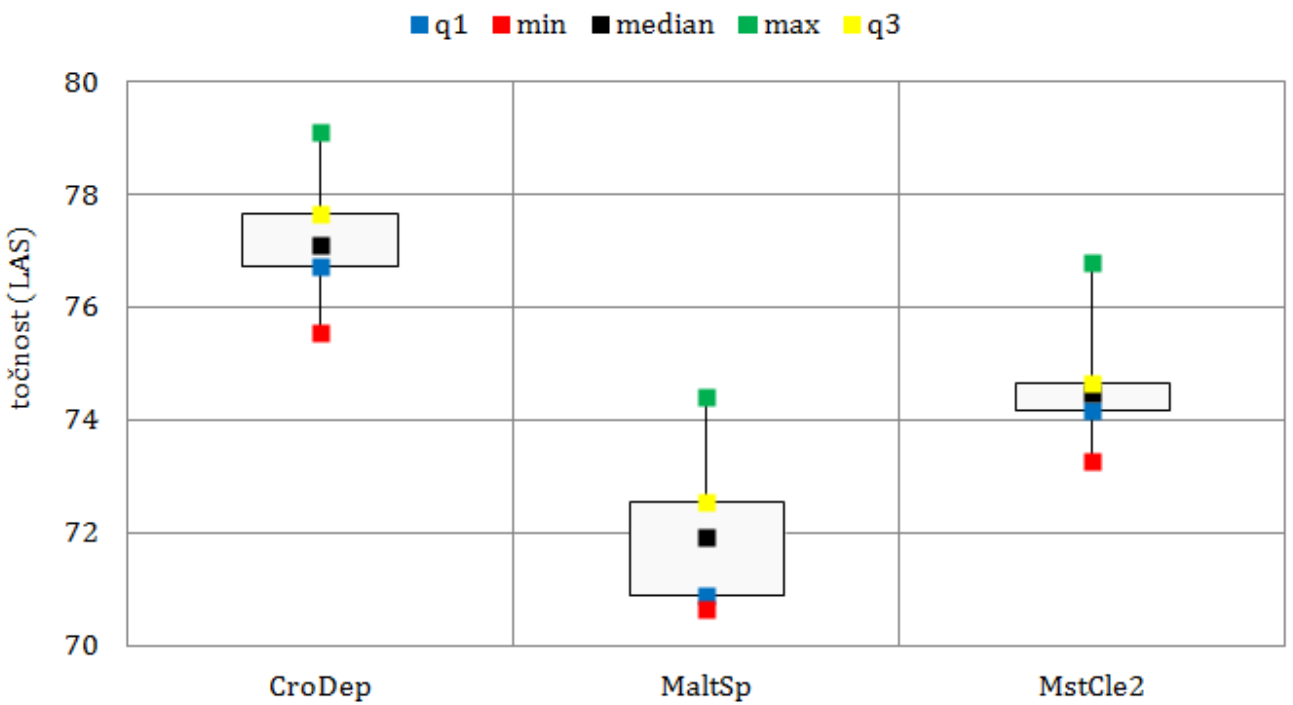
- *k-best* parser s uporabom CROVALLEX-a vrjednovan prema postavkama prethodnoga eksperimenta
- CLE parseru izmjerena samo ukupna točnost prema mjeri LAS u usporedbi s najboljim prijelazničkim parserom i najboljim parserom temeljenim na teoriji grafova prema prethodnome eksperimentu



Predloženi model

Rezultati eksperimenta — ukupna točnost (LAS) i točnost prema vrsti riječi

mjera	N	V	Z	A	S	C	P	R	ukupno
LA	**85.34**	**87.89**	**91.20**	92.67	**98.64**	87.12	**84.38**	80.14	**88.27 ± 0.30**
LAS	**80.10**	**82.85**	73.48	86.40	71.20	**63.24**	76.04	65.77	**77.21 ± 0.59**
UAS	**90.16**	**86.84**	**75.73**	89.13	71.92	**67.06**	84.84	75.30	**83.05 ± 0.50**



Predloženi model

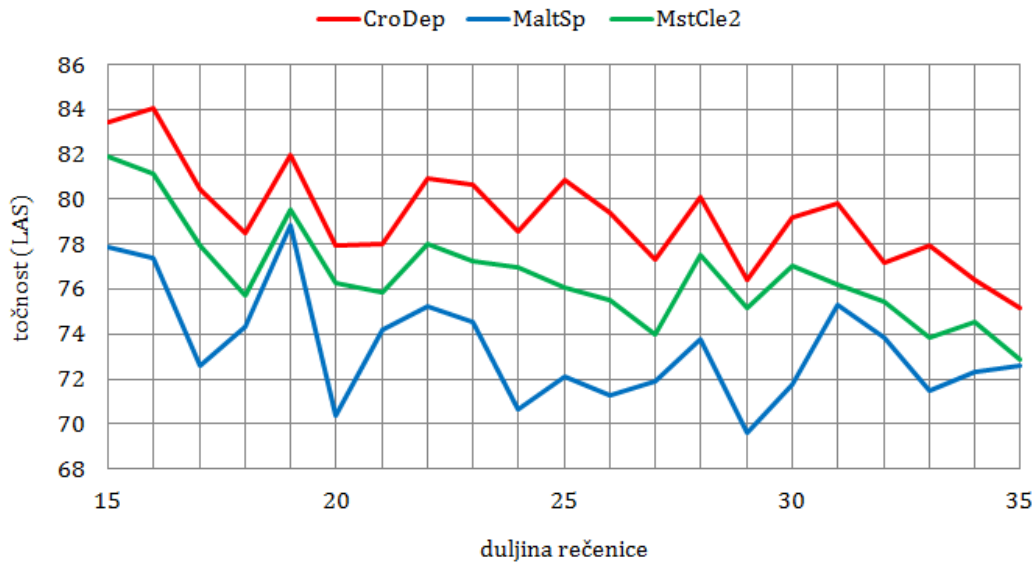
Rezultati eksperimenta — točnost (LAS) s obzirom na sintaktičku funkciju

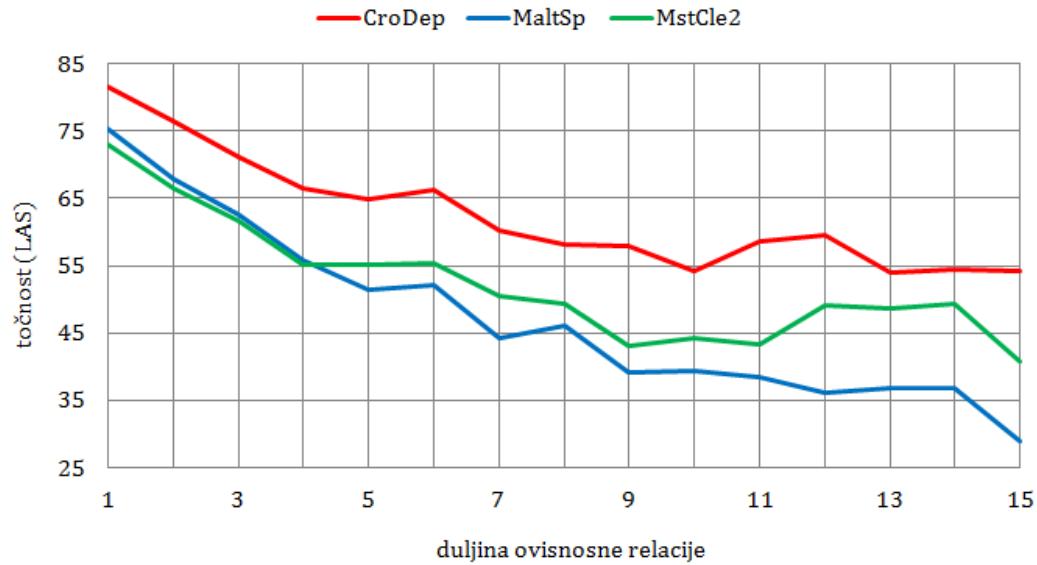
- porast od cca 2.68 LAS u usporedbi s najboljim postojećim sustavom
- porast od najmanje 10.00 LAS za glagole i imenice, odnosno predikate, subjekte i objekte

mjera	Adv	Apos	Atr	AuxC	AuxP	Coord	Obj	Pnom	Pred	Sb
LAS	70.69	34.49	83.94	69.80	70.59	49.41	**83.17**	**71.46**	**82.12**	**85.01**
UAS	84.81	40.99	**88.90**	71.53	71.48	50.87	**93.12**	**79.92**	**86.81**	**91.35**
P (LA)	78.96	58.92	**91.21**	91.96	97.86	89.72	**84.12**	**77.06**	**84.36**	**86.78**
O (LA)	74.11	41.59	90.94	87.77	97.74	81.60	**94.75**	49.73	**97.21**	**97.50**

Predloženi model

Rezultati eksperimenta — LAS s obzirom na duljinu rečenica i ovisnosnih relacija





Predloženi model

Rezultati eksperimenta — vremenska i memorijska učinkovitost

- Intel Core 2 Quad Q6600 (2.40 GHz, 8 MB cache, 1066 MHz FSB), 6 GB radne memorije (DDR2, 1066 MHz)
- Malt* parseri su prijelaznički i parsaju u linearnom vremenu
- prikazano vrijeme parsanja predstavlja zbroj vremena učitavanja modela i njegove primjene
- uporaba CROVALLEX-a ne umanjuje učinkovitost parsera CroDep

postupak	Mjera	CroDep	MaltSp	MstCle2
treniranje	min	**137.79 ± 3.26**	143.9 ± 2.85	328.07 ± 12.16
	MB	~ 2300	**~ 1800**	~ 2800
testiranje	sec	351.74 ± 4.22	470.56 ± 12.11	**143.25 ± 2.18**
	MB	~ 1850	**~ 750**	~ 2200

Zaključak

- iz uvoda
  - Tekstovi pisani hrvatskim jezikom mogu se robustno, točno i učinkovito ovisnosno parsati.
    - najbolji sustav postigao na HOBS-u cca 74.53 LAS
    - parseri temeljeni na teoriji grafova bolji od prijelazničkih parsera
  - Točnost ovisnosnoga parsanja može se povećati uporabom jezičnih resursa za hrvatski jezik, bez gubitka robusnosti i učinkovitosti.
    - uporabljen CROVALLEX i *k-best* ovisnosni parser
    - postignuta točnost od cca 77.21 LAS (povećanje od cca 2.68 LAS)
    - preko 10-postotno uvećanje točnosti za obavijesno najvažnije elemente
    - nema gubitka učinkovitosti

Nacrt budućih istraživanja

- u tijeku
  - utjecaj točnosti lematizacije i MSD-označavanja na točnost ovisnosnog parsanja
    - bitno za uporabu parsera u stvarnim sustavima
    - utjecaj točnosti MSD-označavanja značajniji, posebno u usporedbi s utjecajem točnosti lematizacije
  - utjecaj promjene faktora k na točnost i učinkovitost
    - nema statistički značajnijeg povećanja točnosti
    - gubitak učinkovitosti
  - glasovanje i vezivanje ovisnosnih parsera
    - u tijeku eksperiment s glasovanjem — Malt*, MST* i CroDep
  - uporaba predložene hibridne metode u parsanju drugih jezika
- planirana istraživanja
  - uporaba izostavljenih ovisnosnih parsera
  - uporaba valencijskoga rječnika CROVALLEX u prijelazničkom parsanju
- daljnji razvoj HOBS-a
- lingvistički usmjerenije vrjednovanje točnosti

Hvala na pozornosti! ☺