Skladenjska analiza (ne)standardne slovenščine

Špela Arhar Holdt

Zavod za uporabno slovenistiko Trojina
Filozofska fakulteta Univerze v Ljubljani

Poletna šola Janes, 8. julij 2016

Naloga za današnje predavanje

- Spoznamo korpusni pristop k skladenjskim vprašanjem, metode in vire, ki so za slovenščino trenutno na voljo.
- Ogledamo si nekaj izbranih skladenjskih korpusnih raziskav: zakaj je bila izbrana določena tema, kako je bila zastavljena metodologija – kar je uporabno za morebitno izbiro in zasnovo magistrskih nalog.
- Skupaj razmislimo o uporabnosti skladenjskih podatkov za različne vrste nalog (jezikoslovne raziskave, razvoj jezikovnih tehnologij, prevajanje, lektoriranje ipd.)
- Vaša naloga: med predavanjem si **zapišite eno skladenjsko idejo/ problem/vprašanje,** ki se vam zdi zanimiva za korpusno raziskavo. Zadnji del predavanja bo namenjen debati o zbranih temah.

Skladnja

- „Skladnja ali sintaksa je poseben del slovničnega nauka o jeziku. Uči, kako se delajo (oz. kako so narejene) pravilne povedi in njihovi deli. Povedi in njihove dele skladamo in prepoznavamo na podlagi skladenjskih vzorcev iz besed in stalnih besednih zvez (in iz delov, ki so že sami lahko povedi), pri tem pa zmeraj upoštevamo tudi položaj povedi v danem besedilu oz. v njegovem delu.“ (Toporišič 2004: 487)
- V korpusnem jezikoslovju analiziramo velike količine realnega jezika, da identificiramo in opišemo tipične (in atipične) jezikovne vzorce in značilnosti njihove rabe v danem kontekstu. Zelo poenostavljeno rečeno so skladenjske analize tiste, ki se ukvarjajo z vprašanjem, kako se besede sestavljajo v večje enote (besedne zveze, stavke, povedi, besedila), pri čemer nas zanima tako oblikovna kot pomenska raven.

Tipične skladenjske teme

- Besedne zveze: katere se pojavljajo v rabi, kakšne so njihove oblikovne in pomenske lastnosti ...
- Stavčni členi: kako se pojavljajo v različnih vrstah stavkov, v kakšna razmerja stopajo, kako se izražajo pomenske sestavine ...
- Tipologija stavkov in povedi: katere vrste stavkov poznamo, kako so notranje sestavljeni, kako se povezujejo v širše enote (oblikovno in pomensko) ...
- Raziskave značilnosti upovedovanja: raba trpnika/tvornika; različne vrste naklona; zanikanje; stopnjevanje; nanašanje; izražanje poročanega govora; besedni red ...
- Specifike skladnje v različnih besedilnih vrstah/žanrih: znanstvena besedila, leposlovje ipd. v primerjavi s ‚splošnim jezikom' / primerjava pisnega in govorjenega jezika / primerjava standardne in nestandardne skladnje ...
- ...

Označenost korpusnih podatkov

- Korpusne analize so kvantitativne in kvalitativne. Bistveni predpogoj za oboje je kvalitetna označenost podatkov.
- Označevanje poteka po ravninah: segmentacija, tokenizacija, lematizacija, oblikoskladnja, skladnja, pomenske lastnosti (Polona bo govorila o SRL); + druge raziskovalnospecifične kategorije.
- Večina slovenskih korpusov je označenih do vključno ravni oblikoskladnje, konkordančniki so zasnovani tako, da je po teh podatkih mogoče zelo dobro iskati in jih uporabljati.
- Največji skladenjsko označen korpus je korpus ssj500k, orodja za analizo obstajajo, manjka možnost sintetičnih primerjav (kvantitativni del).
- Določene skladenjske raziskave je mogoče opraviti tudi na neoznačenih podatkih (npr. distribucija členka *pa* v različnih žanrih).

Oblikoskladnja

- Vsaka beseda ima pripisano besedno vrsto in pripadajoče kategorialne lastnosti. Večina večjih korpusov (npr. Gigafida, Kres, GOS, Šolar, Lektor, Janes) je označenih po sistemu JOS: http://nl.ijs.si/jos/.
- Oblikoskadenjske oznake so temelj za večino korpusnih analiz, zato je pomembno, da **sistem označevanja dobro poznamo** – katere oznake so na voljo, kako jih program pripisuje, kje so močne in šibke točke označevanja, kje so razlike med označevanjem in jezikoslovnimi kategorijami, na katerih sistem stoji.
- Dobro razumevanje označevanja je temelj korpusne metodologije (in ključni del vašega metodološkega razdelka).
- Kako torej spoznamo sistem označevanja? Pregledamo oznake sistema in specifikacije, seznanimo se s temeljno literaturo o označevanju konkretnega korpusa, preizkušamo različne iskalne pogoje, sumničavi smo do rezultatov. Včasih pomaga preveriti svoje predpostavke pri kom, ki je sodeloval pri korpusni gradnji.

mesto v kodi	atribut	vrednost	koda
0	besedna_vrsta	pridevnik	P
1	vrsta	splošni	p
		svojilni	s
		deležniški	d
2	stopnja	nedoločeno	n
		primernik	p
		presežnik	s
3	spol	moški	m
		ženski	z
		srednji	s
4	število	ednina	e
		dvojina	d
		množina	m
5	sklon	imenovalnik	i
		rodilnik	r
		dajalnik	d
		tožilnik	t
		mestnik	m
		orodnik	o
6	določnost	ne	n
		da	d

Oblikoskladnja: pridevnik po sistemu JOS

Čeprav je perunika na pogled bolj malo podobna beli liliji, si s to cvetlico deli kar nekaj skupnih simbolnih pomenov. Tako je razkošni modri cvet perunike, ki je bil priljubljen že v starem Egiptu, simbol vladarskega dostojanstva, odličnosti, zmagoslavja in slave. *Gigafida < Gea, 2001*

- beli = [Psnzem] - pridevnik, splošni, nedoločena stopnja, ženski spol, ednina, mestnik
- modri = [Psnmeid] - pridevnik, splošni, nedoločena stopnja, moški spol, ednina, imenovalnik, določna oblika

Primer raziskave 1

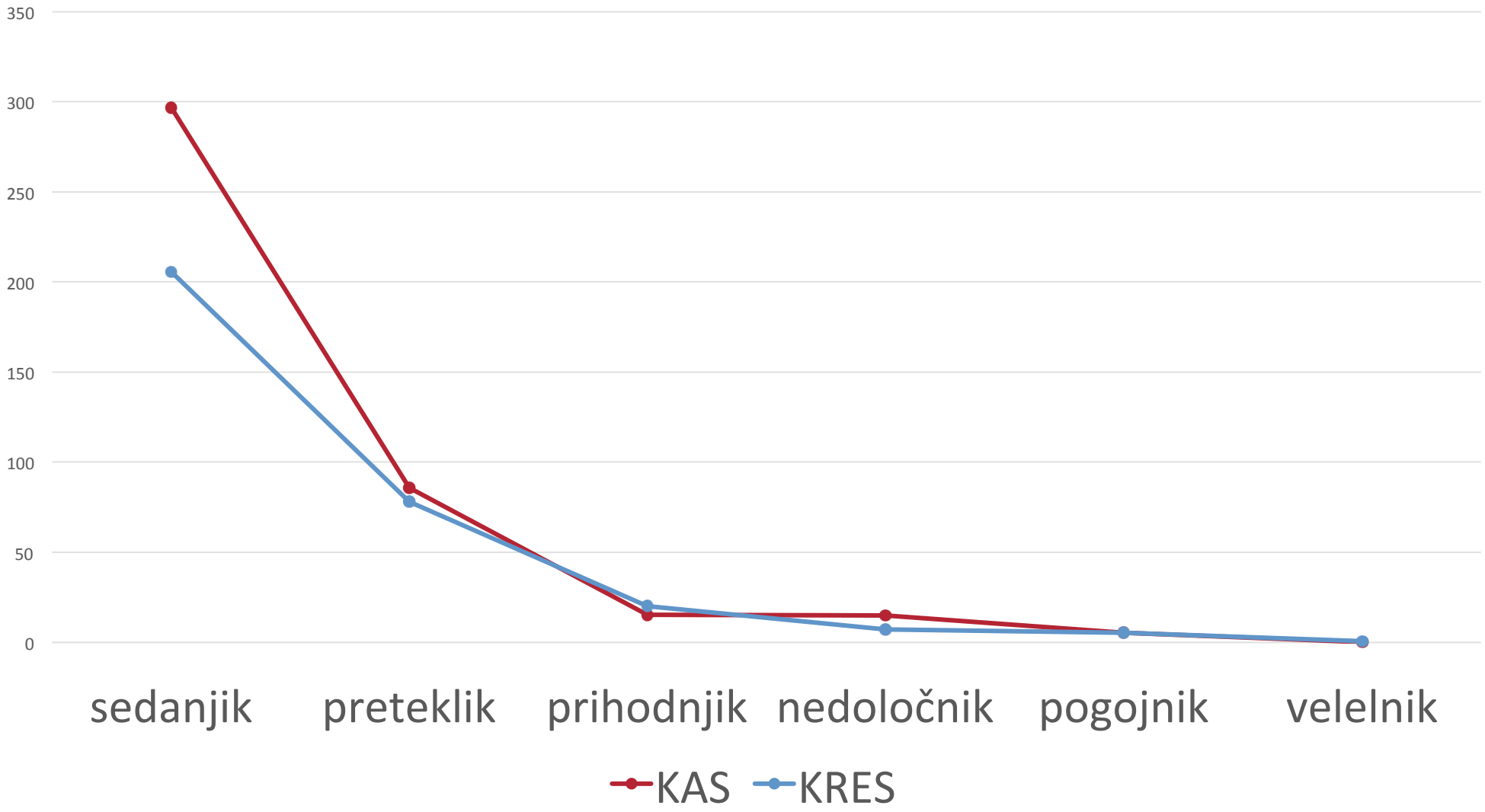
Špela Arhar Holdt, Nataša Logar in Tomaž Erjavec: Slovenska znanstvena besedila: korpusni opis - trpnik. *Toporišičeva Obdobja,* Ljubljana, 10.–12. november 2016. V pripravi.

- Ideja: Ko se v literaturi omenjajo značilnosti znanstvenega jezika, se običajno omenja raba trpnika. Raba trpnika se pogosto tudi vrednoti kot pretirana in slogovno slaba. Vendar te predpostavke niso empirično potrjene.
- Metoda: Izdelamo seznam trpniških skladenjskih vzorcev, ki jih je mogoče pridobiti z uporabo oblikoskladenjskih oznak. Z uporabo konkordančnika SketchEngine primerjamo vzorce v dveh korpusih: referenčnem korpusu Kres in korpusu znanstvenih besedil Kas.
- Rezultat: Podatki pokažejo, da se trpniški vzorci pojavljajo v znanstvenih besedilih pogosteje samo v sedanjiku, v ostalih primerih pa ne. Tako distribucija vzorcev kot njihove tipične leksikalne zapolnitve se v različnih časih in naklonih razlikujejo. V ‚splošnem jeziku' je trpnik zelo prisoten, je pa morebiti manj opazen (prim. da je nekaj *razprodano* ali *operacionalizirano*).

Primer raziskave 1

	primer iz slovnice	CQL za korpusno iskanje na atributu »tag«	korpusni primer
sedanjik	spoštovan je	"Gp-s.*" "Pdn.*"	je prikazan
preteklik	spoštovan je bil	"Pdn.*" "Gp-s.*"	namenjen je (bil)
		"Gp-s.*" "Gp-d.*" "Pdn.*"	je bil ustanovljen
		"Gp-d.*" "Gp-s.*" "Pdn.*"	bil je prepričan
prihodnjik	spoštovan bo	"Gp-p.*" "Pdn.*"	bo uporabljen
		"Pdn.*" "Gp-p.*"	predstavljen bo
nedoločnik	spoštovan biti	"Gp-n.*" "Pdn.*"	biti izpolnjen
		"Pdn.*" "Gp-n.*"	pripravljen biti
pogojnik	spoštovan bi bil	"Pdn.*" "Gp-g" "Gp-d.*"	zaželen bi bil
		"Gp-g" "Gp-d.*" "Pdn.*"	bi bil namenjen
		"Gp-d.*" "Gp-g" "Pdn.*"	bil bi pripravljen
velelnik	bodi spoštovan	"Gp-v.*" "Pdn.*"	bodi pripravljen
		"Pdn.*" "Gp-v.*"	hvaljen bodi

Primer raziskave 1



Primer raziskave 1

	Kas	Kres
najpogostejše zapolnitve	ustanovljen, sprejet, izveden, uporabljen, opravljen, vključen, ugotovljen, namenjen, narejen, določen, objavljen, pripravljen, izbran, dosežen, postavljen	prepričan, sprejet, pripravljen, namenjen, ustanovljen, objavljen, navdušen, izbran, opravljen, rojen, zgrajen, presenečen, izdan, znan, narejen
visoko na seznamu pri obeh korpusih	ustanovljen, opravljen, pripravljen, objavljen, izbran, narejen, izveden, ugotovljen, določen, izdan, zgrajen, postavljen, vključen, prepričan, zaposlen	
primeri, znatno pogostejši v enem korpusu	zajet, zabeležen, predložen, opredeljen, viden, zastavljen, dodan, omogočen, obdelan, poudarjen, dopolnjen, izločen, realiziran, osredotočen, pregledan	navdušen, ranjen, presenečen, ukraden, narisan, naperjen, poražen, posiljen, odrezan, obarvan, prepleten, dvignjen, najavljen, razbit, spočet
korpusnospecifični primeri	optimiziran, dimenzioniran, odčitan, izpodbit, reorganiziran, kategoriziran, podcenjen, operacionaliziran, poiskan, oskrbovan, zmodeliran, pozicioniran	opustel, preklan, kostumiran, otrpel, zaljubljen, izžreban, upadel, razprodan, opevan, razkačen, očitan, pridušen, negovan, deponiran, izmozgan

Skladenjsko označevanje

- „Rezultat označevanja so skladenjsko označeni korpusi oz. drevesnice, ki predpostavljena skladenjska razmerja eksplicirajo na velikem vzorcu besedil dejanske rabe in omogočajo statistični pregled vzorcev distribucije skladenjskih struktur, zato so navadno izhodišče za nadaljnjo, bolj poglobljeno jezikoslovno analizo, hkrati pa so pomembni za razvoj jezikovnih tehnologij, kar je temeljnega pomena za ohranjanje konkurenčnosti ter polnofunkcionalnosti jezika med ostalimi jeziki“ (Ledinek 2014: 18).
- Skladenjsko označevanje (ročno) vs. razčlenjevanje (strojno).
- Sistemi za skladenjsko označevanje so zelo različni, od plitkih oz. skeletnih modelov (zamejitev stavčnih enot z določitvijo jedra) do popolnega oz. globinskega označevanja, kjer označujemo funcijskoskladenjska in pomenskoskladenjska razmerja med elementi povedi (ibid: 19).

Skladenjsko označevanje

- Odvisnostni sistemi temeljijo na odnosih med elementi (kjer se predideva, da je eden od delov jedrni, drugi od njega odvisen). So običajna izbira za morfološko bogate jezike s prostejšim besednih redom. Drugi glavni princip je frazna gramatika.
- Za slovenščino: Slovenska odvisnostna drevesnica po vzoru Prague Dependency Treebank (Erjavec in Ledinek 2006); korpus ssj500k, ki je skladenjsko označen z odvisnostnim sistemom JOS-SSJ (Dobrovoljc et al. 2012) in je bil pred kratkim tudi preveden v model Universal Dependencies (Dobrovoljc et al. 2016).

Ssj500k

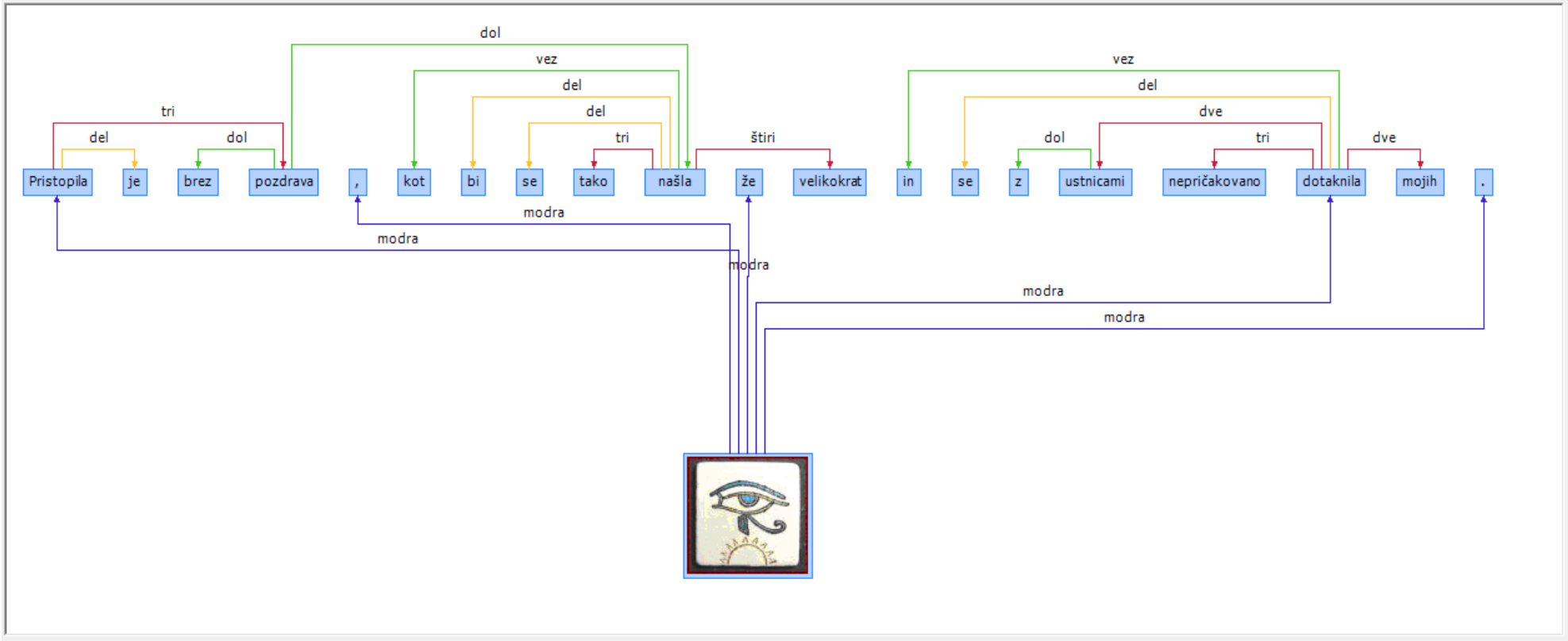
- Korpus obsega 11.411 skladenjsko označenih povedi (Krek et al. 2015). Ob pripravi podatkov je bil razvit razčlenjevalnik, ki ga lahko preizkusimo na http://razclenjevalnik.slovenscina.eu/.
- Podatke ssj500k lahko pregledujemo v označevalniku SSJ („Brank“), ki omogoča tudi preoznačevanje (torej je uporaben tako za analize obstoječega gradiva kot za označevanje novega).
- Tudi pri skladenjski ravni je za kvalitetno izvedbo raziskav treba dobro poznati sistem označevanja.
- Na korpusu ssj500k še ni bilo opravljenih veliko raziskav. Pomembna izjema je monografija N. Ledinek (2012), ki prinaša študijo slovnično-pomenskih vzorcev za 20 slovenskih glagolov.
- Korpus, označevalnik in razčlenjevalnik so prosto dostopni za uporabo (www.slovenscina.eu).

Skupina povezav	Tip povezave	Kaj povezuje
Povezave **prvega nivoja** označujejo razmerja znotraj besednih zvez.	*dol*	Jedro in določilo besednih zvez.
	del	Deli zloženega povedka.
	prir	Jedra v prirednih zvezah znotraj stavka.
	vez	Besede ali ločila v vezniški vlogi.
	skup	Nepolnopomenske besede,ki imajo zelo močno tendenco po sopojavljanju.
Povezave **drugega nivoja** označujejo stavčne člene.	*ena*	Osebek stavka.
	dve	Predmet stavka.
	tri	Prislovno določilo lastnosti.
	štiri	Ostala prislovna določila.
Povezava **tretjega nivoja** se uporablja za povezovanje vseh ostalih struktur.	*modra*	Hierarhično najvišje pojavnice, skladenjsko manj predvidljive in oddaljene strukture, vrinki, ločila.

C:\Users\Špela\OneDrive\Documents\#JANES\#OZNAČEVANJE\ssj500k.xml - Označevalnik stavkov

Izberi datoteko... | Prikaži vse stavke | | Išči | Povezave... | Urejanje izbranega stavka | Nastavitve...

[ssj5.19.67] Lahko si nekaj mislimo .
[ssj5.20.68] Oblekla si je bila kavbojke in pleten pulover , lase si je spletla v rep , ki ji je ljubko štrlel na zatilju ; bila je drugačna , kot sem je bil vajen , zdela se mi je nekako veliko bolj vsakdanja in za to sem ji bil hvaležen .
[ssj5.20.69] Pristopila je brez pozdrava , kot bi se tako našla že velikokrat in se z ustnicami nepričakovano dotaknila mojih .
[ssj5.20.70] Želela je najverjetneje , da bi bil to le blag dotik ob srečanju , toda moje dlani so se oklenile njenih gladkih lic in je niso več spustile .
[ssj5.20.71] Čutil sem njene polne nenašminkane ustnice , ki so polagoma dojemale , kako zelo si jih želim .
[ssj5.20.72] Vohkal sem njen topli , najverjetneje še tudi od hitre hoje razgreti dih , in se prepuščal občutku lepega .
[ssj5.20.73] Spoznaval sem , morda prvič , da o vsej silni želji in hotenju , ki ga je zganila v meni , ni mogoče povedati z besedami , zato nisem niti poskusil .
[ssj5.21.74] Na voljo naj bi imeli blizu milijardo tolarjev (10 milijonov DEM) več kot pred petimi leti
[ssj5.22.75] V nedeljo , 23. maja , bo v Križankah v Ljubljani nastopil Julio Iglesias - Cene vstopnic za koncert se gibljejo od 19 do 29 tisoč tolarjev , kar za slovenski žep nikakor ni malo - Zato smo mimoidoče Ljubljančane povprašali , a
[ssj5.23.76] V osnovni šoli v Bistrici pri Tržiču bodo v soboto , 8. , in nedeljo , 9. maja , pripravili 27. Mednarodne dneve mineralov , fosilov in okolja .
[ssj5.23.77] Predstavilo se bo sto razstavljavcev iz petnajstih držav , ki bodo razstavljali in prodajali naravne in obdelane lepote skritega podzemlja .
[ssj5.23.78] Razstavo bo v petek , 7. maja , ob 19. uri odprl Ivo Bizjak , varuh človekovih pravic .
[ssj5.24.79] Tako je bil včeraj obiskan profesionalni teniški turnir ATP v Domžalah , ko je nastopil Mariborčan Iztok Božič .
[ssj5.25.80] Kot smo že poročali , je bila afera Holmec eden od razlogov , da so v parlamentu zamenjali notranjega ministra Mirka Bandlja (LDS) .
[ssj5.25.81] Na predlog premiera Janeza Drnovška so za naslednika na vrhu notranjega ministrstva imenovali Boruta Šukljeta .

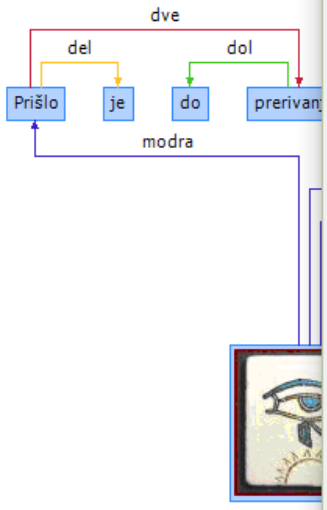


C:\Users\Špela\OneDrive\Documents\#JANES\#OZNAČEVANJE\ssj500k.xml - Označevalnik stavkov

Izberi datoteko... | Prikaži vse stavke | Išči | Povezave... | Urejanje izbranega stavka | Nastavitve...

[ssj1.1.3] Dekle je ob vzvratni vožnji začelo vpiti , da bi jo utišal , sem prijel nož .
[ssj1.1.4] Prišlo je do prerivanja in umrla je .
[ssj1.1.5] Tega se sploh nisem zavedel .
[ssj1.1.6] Kaj se je zgodilo , sem izvedel šele naslednjega dne iz časopisja , " je v intervju za Stampo iz Torina izjavil morilec .
[ssj1.1.7] Preiskave med sodnim postopkom so pokazale , da so se dogodki odvijali bistveno drugače .
[ssj1.1.8] Dogodki v prihodnjih mesecih pa bodo pokazali , ali bo morilcu iz Ankarana tokrat uspelo prepričati italijanske pravosodne oblasti .
[ssj2.2.9] V bolnišnici bodo uvedli tudi s š
[ssj2.2.10] V bolnišnici so že pred časom
[ssj2.2.11] Začetek izvajanja programa je
[ssj2.3.12] Piran - Čeprav je vodstvu pira
[ssj2.3.13] Zato je včeraj sklicala sestanel
[ssj3.4.14] Kar zadeva podočnjake , je st
[ssj3.4.15] Lahko so posledica vnetja vek
[ssj3.4.16] Najprej je treba najti vzrok , z
[ssj3.4.17] Včasih so podočnjaki tudi posle

o prihodnji mesec .
var države , se je ob zadnjih zapletih županja Vojka Štular odločila , da jim



Iskanje povezav

	A	B	C	D
MSD	G*	D*	S*	
Lema				
Oblika				

Išči | Briši | Shrani...

Pogoji za povezave

Briši | obstaja | povezava od A | do C | tipa dve
Briši | obstaja | povezava od C | do B | tipa dol
Dodaj

Stavek	A	B	C	A -> C	C -> B
ssj1.1.4	Prišlo	do	prerivanja	dve	dol
ssj2.3.13	prišli	do	rešitve	dve	dol
ssj3.4.15	trpijo	za	alergijami	dve	dol
ssj3.6.19	govorimo	o	motnji	dve	dol
ssj3.6.19	izmenjavajo	z	obdobji	dve	dol
ssj3.7.24	Opozorila	na	komunikacijo	dve	dol
ssj3.7.24	gre	za	zdravljenje	dve	dol
ssj3.8.30	upravlja	z	aparatom	dve	dol
ssj3.9.32	zaleže	Pri	bolečinah	dve	dol
ssj3.14.47	vplivala	na	okolje	dve	dol
ssj4.15.48	postregla	z	naključji	dve	dol
ssj4.15.50	sovpadal	z	dnevom	dve	dol
ssj4.15.51	križal	z	načrti	dve	dol
ssj4.15.54	prikrajšati	za	prenos	dve	dol
ssj5.19.64	gre	za	diskriminacijo	dve	dol
ssj5.20.69	dotaknila	z	ustnicami	dve	dol
ssj5.20.73	povedati	o	želji	dve	dol
ssj5.25.81	imenovali	za	naslednika	dve	dol
ssj5.26.83	Gre	za	vozilo	dve	dol
ssj5.26.84	primerja	z	avtomobilom	dve	dol

Dvojni klik = prikaz izbrane povezave v glavnem oknu.

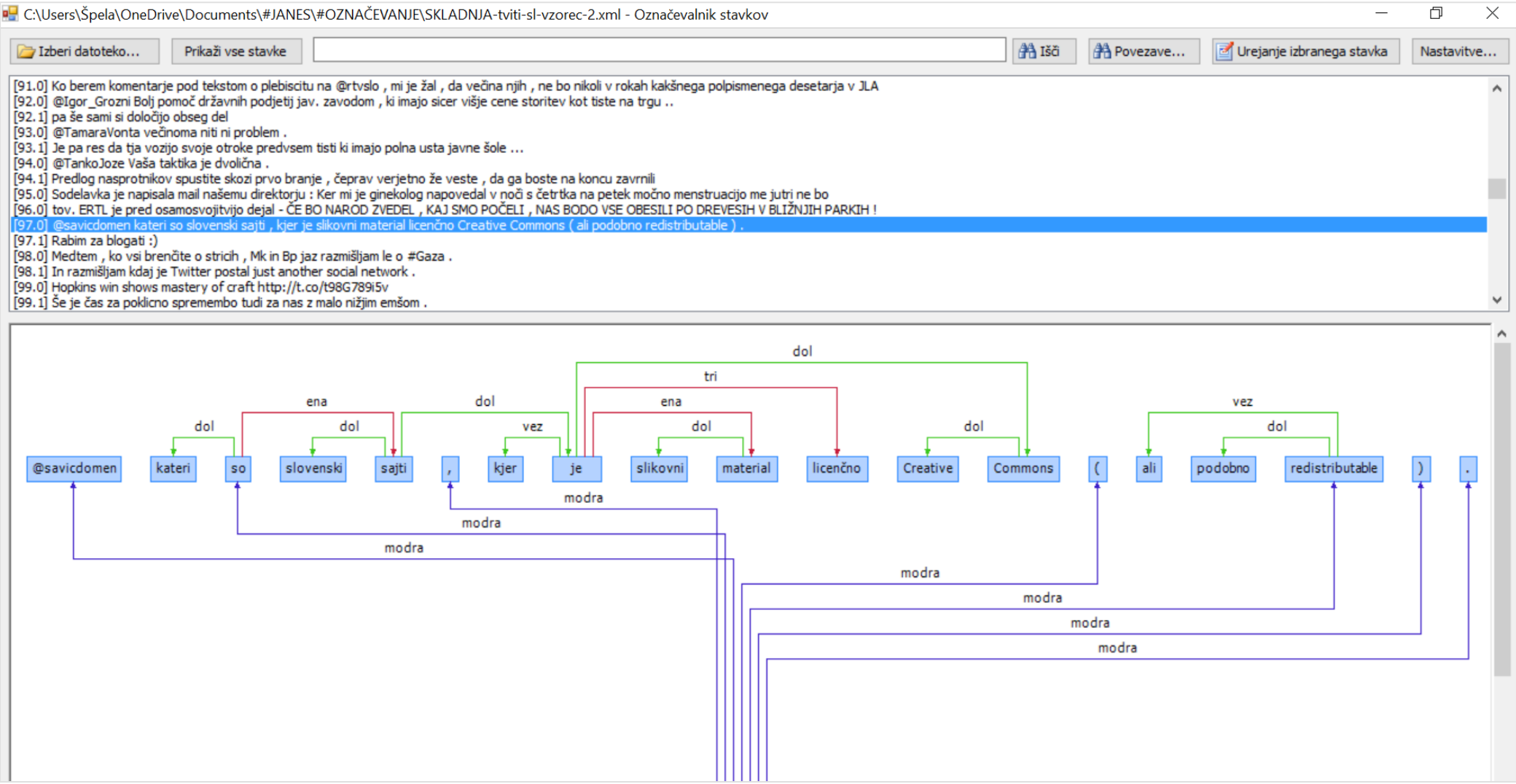
Raziskave nestandardne skladnje

- Nestandardna skladnja se v preteklosti pogosto omenja na liniji prav-narobe oz. s stališča vrednotenja sloga. Primer iz Nove slovenske skladnje (Toporišič 1982):

Ob veznikih *ne samo/le — ampak/temveč/marveč tudi* prestava v rodilnik ni obvezna:[30] *Tega/To nisi pravil le meni, ampak vsakemu, ki te je hotel poslušati.* Izjemoma se tako govori še pri *nič*: *Nič/Ničesar ji ne manjka pri meni, Nič/Ničesar mi ne pripoveduj!, Nimam kaj/česa obžalovati.* — Neknjižno je, če postavljamo v rodilnik osebek drugih zanikanih povedkov, npr. *Ničesar (prav nič) se ni zgodilo.* Tudi pri trpniku je prestava osebka v rodilnik nedopustna: *Takih stvari se ne govori na glas* (namesto *Take stvari se ne govorijo na glas*), *Tega se mi ni naročilo* nam. *To se mi ni naročilo.*[31] Napačno je tudi *Tega mi ni bilo naročeno.*[32]

Raziskave nestandardne skladnje

- Prve analize, v katerih se opazuje značilnosti skladnje v računalniško posredovani komunikaciji na večji količini gradiva v Michelizza (2015).
- Širših korpusnih študij še ni.
- S sistemom JOS-SSJ smo označili 200 tvitov, na osnovi katerih bodo dopolnjenje smernice za označevanje.
- S tem delom bi bilo mogoče nadaljevati (označiti več primerov, morda tudi druge besedilne vrste, natrenirati razčlenjevalnik). Če koga zanima, naj se javi!
- Problem raziskovanja nestandardne skladnje: manjka korpusno osnovani slovnični opis standardnega jezika. Če ne vemo, kaj je v jeziku tipično standardno, tudi ne moremo empirično opredeljevati, kaj je nestandardno.


C:\Users\Špela\OneDrive\Documents\#JANES\#OZNAČEVANJE\SKLADNJA-tviti-sl-vzorec-2.xml - Označevalnik stavkov
Izberi datoteko...
Prikaži vse stavke
Išči
Povezave...
Urejanje izbranega stavka
Nastavitve...
[91.0] Ko berem komentarje pod tekstom o plebiscitu na @rtvslo , mi je žal , da večina njih , ne bo nikoli v rokah kakšnega polpismenega desetarja v JLA
[92.0] @Igor_Grozni Bolj pomoč državnih podjetij jav. zavodom , ki imajo sicer višje cene storitev kot tiste na trgu ..
[92.1] pa še sami si določijo obseg del
[93.0] @TamaraVonta večinoma niti ni problem .
[93.1] Je pa res da tja vozijo svoje otroke predvsem tisti ki imajo polna usta javne šole ...
[94.0] @TankoJoze Vaša taktika je dvolična .
[94.1] Predlog nasprotnikov spustite skozi prvo branje , čeprav verjetno že veste , da ga boste na koncu zavrnili
[95.0] Sodelavka je napisala mail našemu direktorju : Ker mi je ginekolog napovedal v noči s četrtka na petek močno menstruacijo me jutri ne bo
[96.0] tov. ERTL je pred osamosvojitvijo dejal - ČE BO NAROD ZVEDEL , KAJ SMO POČELI , NAS BODO VSE OBESILI PO DREVESIH V BLIŽNJIH PARKIH !
[97.0] @savicdomen kateri so slovenski sajti , kjer je slikovni material licenčno Creative Commons (ali podobno redistributable) .
[97.1] Rabim za blogati :)
[98.0] Medtem , ko vsi brenčite o stricih , Mk in Bp jaz razmišljam le o #Gaza .
[98.1] In razmišljam kdaj je Twitter postal just another social network .
[99.0] Hopkins win shows mastery of craft http://t.co/t98G789i5v
[99.1] Še je čas za poklicno spremembo tudi za nas z malo nižjim emšom .
dol
tri
ena
dol
ena
vez
dol
dol
dol
dol
vez
dol
@savicdomen
kateri
so
slovenski
sajti
,
kjer
je
slikovni
material
licenčno
Creative
Commons
(
ali
podobno
redistributable
)
.
modra
modra
modra
modra
modra
modra
modra

Primer – zaznamovan besedni red?

L1T1 (3/48)

- Nekateri zvesti podporniki.. Še vam ni jasno, da če bi želeli videti vsak tweet kandidatov, bi enostavno sledili njim? #predsednik12
- peter_pec Plus, premikanje na SD kartico sem probala že stokrat. Tudi telefon to ponudi kot možnost. Rezultat? "Ni predmetov za premikanje"
- thetide so Ready To Go. Se veselim dobre letine pristankov. #aritmija #koncert #landings #novaplata http://t.co/mytlEbvTUz

Primer – zaznamovan besedni red?

L3T3 (7/49)

- union_pivo pizda, zakaj morm zmer uniona pit z laško kozarca? A to je taka politika al nimate kozarcev al vam je vseeno za kulturo piva??!!
- merineseri pa če ni snicekrsaaaa ;) zdej sm se spomnu, da mamo doma doma arašide, k jih je treba res kopat iz zemlje ;)
- TaiaG ja no, men se to tud skoooz dogaja! Se veckrat pa s senckami na instagramu. Nardim smokey zgleda pa cist neki neznega :/
- Njokifestival ma ja...butaste @SlovenskeNovice so se na @francikek spravle...#kreteni...to je čist legalno... @TinoMamic @ErikaPlaninsec
- AlanBStard2 @MajdaSirca @stanka_d @SoMe_Meli jaz sem posnel na RŠ in prinesel na golf v Kr-je folk čist znoru,danes malo nerodno heh

Primer raziskave 2

Arhar Holdt, Š. in Dobrovoljc, K. (2016): Vrednost korpusa Janes za slovensko normativistiko. *Slovenščina 2.0*, 4 (2): 1–37.

- Ideja: Primerjati, kako se pojavljajo zveze samostalnika z neujemalnim levim prilastkom (*solo petje, RTV prispevek*) v korpusih Janes in Kres.
- Metoda: Z uporabo konkordančnika SketchEngine primerjamo vzorce v dveh korpusih: referenčnem korpusu Kres in korpusu Janes.
- Rezultat: Analiza razkriva: da se referenčni korpus Kres in korpus Janes glede zapisa teh zvez pomembno razlikujeta; da je raba tovrstnih zvez v korpusu Janes pogostejša in bolj raznolika kot v korpusu Kres; da se v obeh korpusih pojavlja visok delež zvez, ki v rabi izkazujejo variantnost v zapisovanju, tudi na ravni posameznih prilastkov; in – vsaj na prvi pogled – presenetljivo, da je raba v korpusu Janes konsistentnejša, kar nakazuje, da jezikovna regulacija obravnavanega problema povečuje variantnost v jezikovni rabi.

Primer raziskave 2

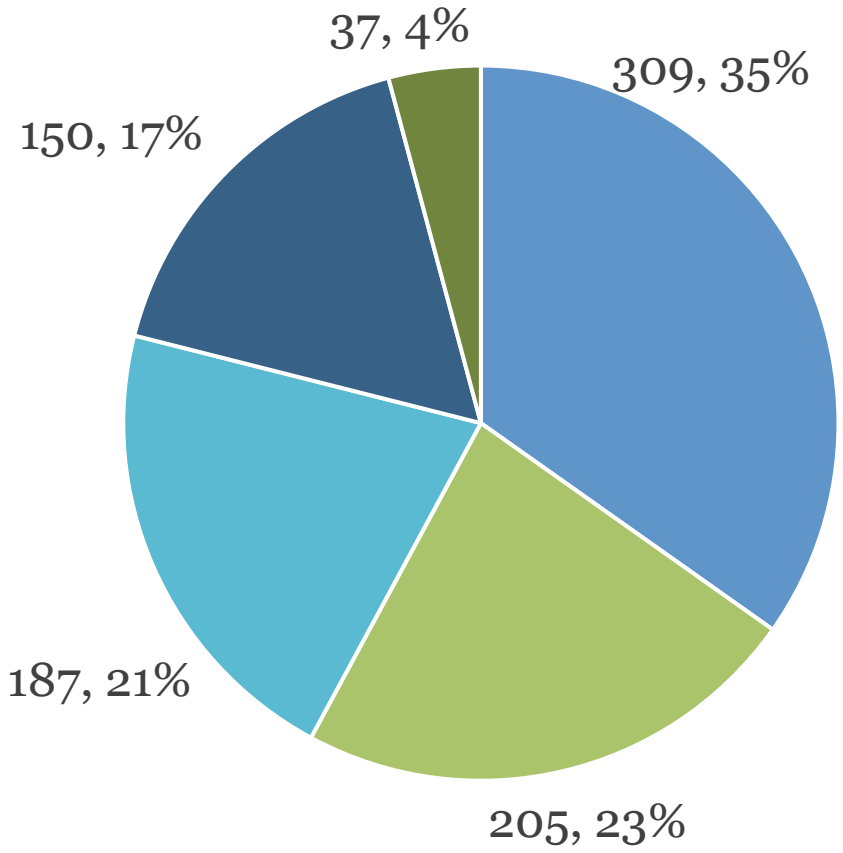
- Kot potencialne zveze samostalnikov z neujemalnim levim samostalniškim prilastkom smo z orodjem noSketchEngine iz obeh korpusov izluščili nize dveh zaporednih samostalnikov, samostalnika v imenovalniku in samostalnika s poljubnimi oblikoskladenjskimi lastnostmi, pri katerih se dana oblika prvega samostalnika ne glede na velikost črk v celotnem korpusu pojavi pred vsaj tremi različnimi oblikami leme jedrnega samostalnika (npr. *rtv prispevek*, *rtv prispevka*, *rtv prispevkom*). Če je bil ta pogoj izpolnjen, je bil niz oblike prilastka in leme jedra prepoznan kot potencialna zveza samostalnika z neujemalnim levim samostalniškim prilastkom (*rtv prispevek*).
- Pregled rezultatov je razkril, da tokenizacijske, besednovrstne in oblikoskladenjske interpretacije, kakršne so bile korpusnim besedilom pripisane v postopku strojnega oblikoskladenjskega označevanja, na podatke vplivajo tako zaradi neenotnosti označevalnikov kot zaradi nedoslednosti v obstoječih jezikovnih virih.

Primer raziskave 2

	Kres		Janes		Skupno
	Abs.	Rel.	Abs.	Rel.	Abs.
pojavnice (vse)	95.897	987	212.808	1.662	
pojavnice (pregibne)	49.047	505	108.303	846	
različnice – zveze	3.054	**31**	7.840	**61**	888
različnica – prilastki	1.432	**15**	2.851	**22**	719

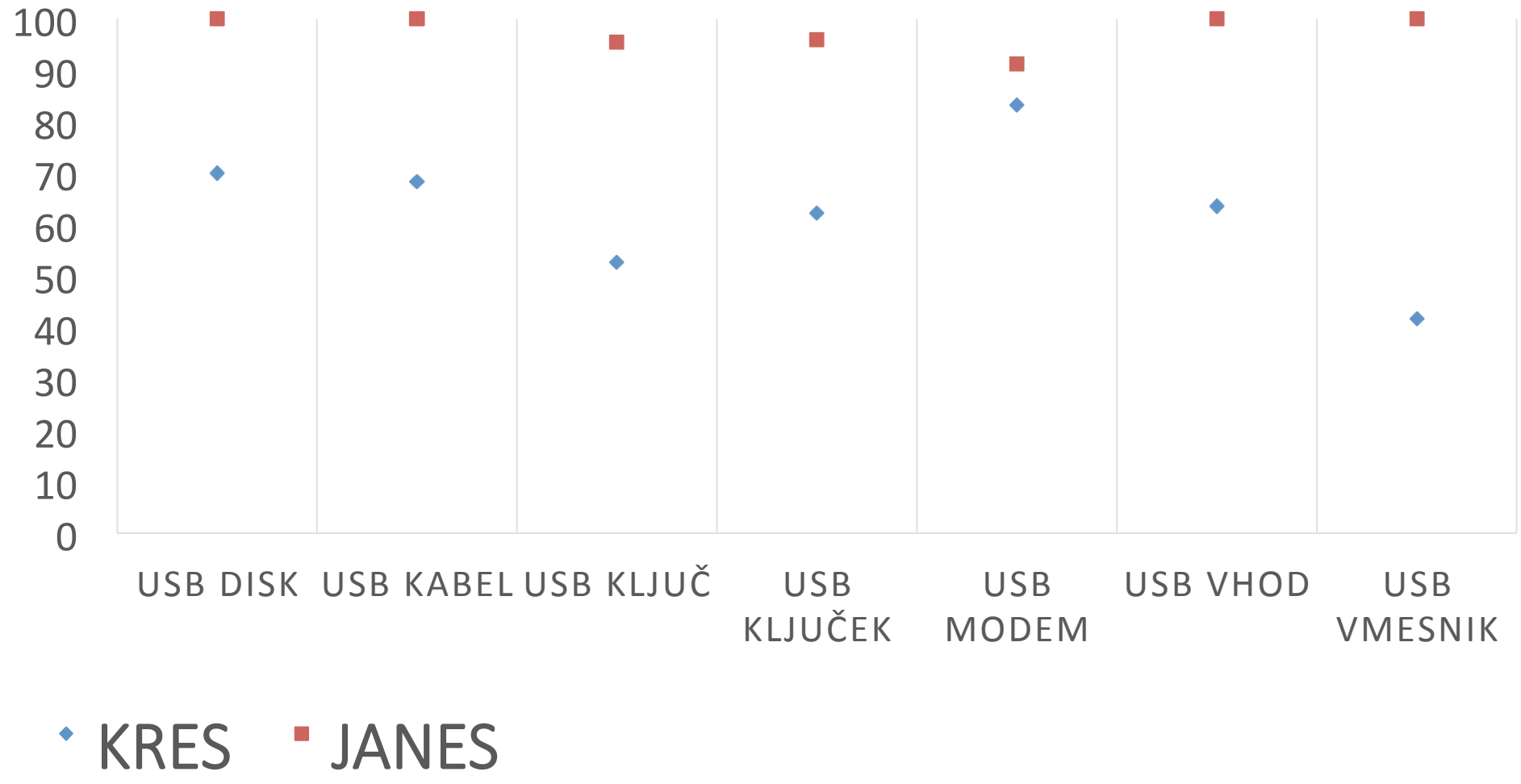
Načini zapisa zveze	Kres	Janes
samo zapis narazen (npr. *loto številka*)	71 %	75 %
zapis narazen in z vezajem (npr. *tv film, tv-film*)	13 %	8 %
zapis narazen in skupaj (npr. *špas teater, špasteater*)	11 %	9 %
zapis narazen, z vezajem in skupaj (npr. *new york, newyork, new-york*)	5 %	7 %

Primer raziskave 2

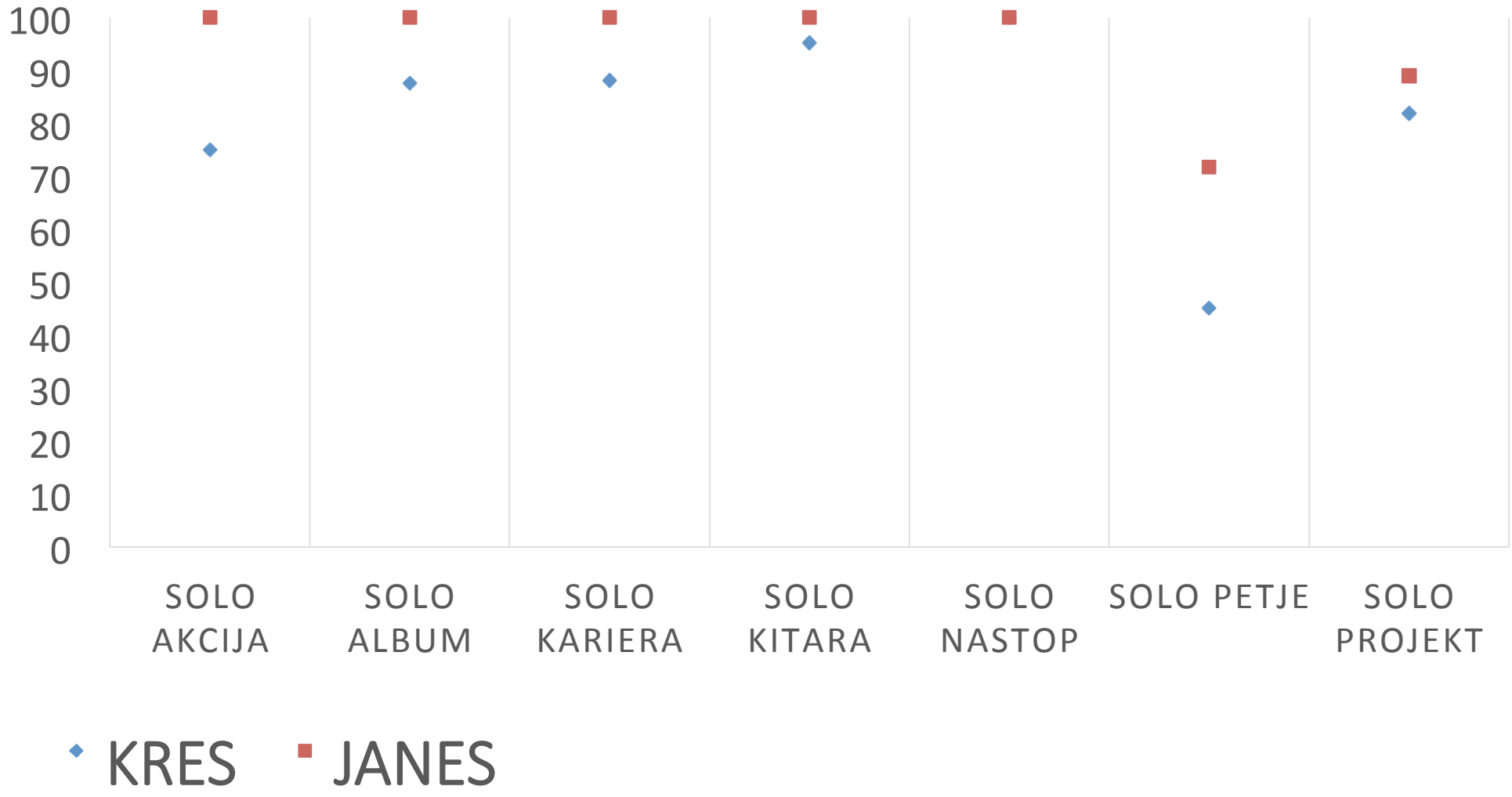


- Zveze z neujemalnim prilastkom (joga studio, android telefon)
- Lastna imena (butan plin, indiana jones)
- Kratične zveze (usb ključek, c vitamin)
- Nerelevantni rezultati (york city, pearl jama)
- Citatna oz. polcitatna imena (after party, bad boy)

Primer raziskave 2



Primer raziskave 2



T'ain't what you do
it's the way
that you do it

that's what gets results.

"T'ain't What You Do (It's the Way That You Do It)" is a song written by jazz musicians Melvin "Sy" Oliver and James "Trummy" Young. It was first recorded in 1939 by Jimmie Lunceford, Harry James, and Ella Fitzgerald.[1]

Hvala za pozornost!

spela.arhar@trojina.si
Spela.ArharHoldt@ff.uni-lj.si

Literatura in nadaljnje branje

- Arhar Holdt, Š., Nataša Logar in Tomaž Erjavec: Slovenska znanstvena besedila: korpusni opis - trpnik. *Toporišičeva Obdobja,* Ljubljana, 10.–12. november 2016. V pripravi.
- Arhar Holdt, Š., Dobrovoljc, K. (2016): Vrednost korpusa Janes za slovensko normativistiko. *Slovenščina 2.0*, 4 (2): 1–37.
- Böhmová, A., Hajič, J., Hajičová, E. and Hladká, B. (2003). The Prague dependency treebank. In *Treebank: Building and Using Parsed Corpora*. Netherlands: Springer, pp. 103–127.
- Crystal, D. (2011). *Internet Linguistics: A Student Guide.* London, New York: Routledge.
- Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '16).* Portorož, pp. 1566–73.
- Dobrovoljc, K., Erjavec, T. and Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. In *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana (in print).
- Dobrovoljc, K., Krek, S. and Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik 8. konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan, pp. 42–47.
- Erjavec, Tomaž in Ledinek, Nina, 2006: Slovenska odvisnostna drevesnica, prvi rezultati. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik 5. konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan, pp. 162–167.
- ERJAVEC, Tomaž, idr., 2008: *Specifikacije za leksikon besednih oblik*: *SSJ kazalnik 3*. Kamnik. http://projekt.slovenscina.eu/Vsebine/Sl/Kazalniki/K3.aspx
- Erjavec, T., Fišer, D., Krek, S. and Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In: *LREC 2010, 7th International Conference on Language Resources and Evaluations*. Valletta, pp. 1806–1809.
- Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S. and Velušček, A. (2008). *Specifikacije za učni korpus.* Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ.
- Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C. and Smith, N. A. (2014). A dependency parser for tweets. In *Proc. of EMNLP*. Doha, Qatar, pp. 1001–1012.
- Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N. and Holz, N. (2015). *Training corpus ssj500k 1.4, Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.
- Ledinek, Nina, 2014: Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine. Ljubljana: Založba ZRC, ZRC SAZU.
- Michelizza, M. (2015). *Spletna besedila in jezik na spletu.* Založba ZRC, ZRC SAZU, Ljubljana.
- Toporišič, Jože, 1982: Nova slovenska skladnja. Ljubljana: DZS.
- TOPORIŠIČ, Jože, 2004: *Slovenska slovnica: četrta, prenovljena in razširjena izdaja, 2. natis.* Maribor: Obzorja.