Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres

ŠPELA ARHAR HOLDT[1,2] IN KAJA DOBROVOLJC[1]

1) ZAVOD ZA UPORABNO SLOVENISTIKO TROJINA

2) FILOZOFSKA FAKULTETA UNIVERZE V LJUBLJANI

Izhodišče raziskave

- O vprašanju zapisovanja/jezikovnosistemskega uvrščanja zvez, kot so ***alfa samec, servo volan, RTV prispevek*** (ali kot medponskoobrazilne zloženke ***alfasamec, servovolan, RTV-prispevek***), se v slovenistiki že dolgo razpravlja. (Logar 2005)
- V zadnjih letih se sistemsko-teoretičnim razpravam pridružujejo korpusne. Ker pa je raba tovrstnih zvez pogosto korigirana, se izkazuje, da na osnovi referenčnih korpusov (ki vsebujejo velik, vendar ne natančno določljiv delež lektoriranih besedil) ni mogoče realno oceniti obsežnosti in narave obravnavanega problema. (Dobrovoljc in Jakop 2011: 113–114; Logar 2012)
- Novonastali korpus Janes ponuja možnost za vpogled v jezikovno produkcijo brez lektorskih oz. uredniških posegov (ob upoštevanju specifik v korpus zajetih besedilnih vrst in izogibu vprašanj o samokorekciji).
- Hipoteza: glede pojavljanja in rabe obravnavane vrste besed oz. zvez se **uravnoteženi referenčni korpus Kres** (Logar et al. 2012) in **korpus uporabniških vsebin Janes** (Fišer et al. 2015) pomembno razlikujeta.
- Namen prispevka: ugotoviti razlike glede pojavljanja in rabe v obeh korpusih in s tem preveriti vrednost korpusa Janes za normativistične raziskave.

Luščenje korpusnih podatkov

- Programska skripta, ki v prvem koraku lušči zveze v zapisu narazen (***alfa samec, servo volan, RTV prispevek***).
- Iz obeh korpusov sva izluščili nize **dveh zaporednih samostalnikov (S.* S.*),** pri katerih se nespremenjena oblika prvega samostalnika pojavi pred **vsaj tremi različnimi oblikami** drugega samostalnika.

  servo volan, servo volana, servo volanom → *servo volan*
- **Problem 1: Različnost označevanja.** Korpus Kres je označen z označevalnikom Obeliks (Grčar et al. 2012) in korpus Janes z označevalnikom ToTaLe (Erjavec et al. 2005). Čeprav oba označevalnika svoj model znanja gradita na istih jezikovnih virih, leksikonu besednih oblik Sloleks (Dobrovoljc et al. 2015) in učnem korpusu ssj500k (Krek et al. 2013), prihaja med njima do določenih razlik, ki otežujejo primerjavo korpusnih podatkov.
- **Problem 2: Nedosledna obravnava nesklonljivih prilastkov v jezikovnih virih,** zlasti pri vprašanjih besednovrstne kategorizacije (kako ločujemo med pridevniki in samostalniki) in njihove obravnave v besedilnem kontekstu (kako slovnične lastnosti jedra vplivajo na označevanje spola, sklona in števila nesklonljivih pridevnikov oz. sklona nesklonljivih samostalnikov v vlogi pridevnika. (Gl. tudi Gantar 2015)

K problemom se vračamo na koncu predstavitve.

Izluščeni kandidati

	Kres različnice		Kres pojavnice		Janes različnice		Janes pojavnice		Prekrivne različnice
Pogostost	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.
zveze (npr. *alfa samec*)	3.054	**31**	95.897	987	7.840	**61**	212.808	1.662	888
prilastki (npr. *alfa*)	1.432	15	95.897	987	2.851	22	212.808	1.662	719

- Že hiter pregled številčnih podatkov potrdi hipotezo o pomembni različnosti korpusnega gradiva: zveze kandidatke so v korpusu Janes skoraj enkrat **pogostejše** kot v korpusu Kres, obenem pa so na ravni prilastkov tudi bolj **raznolike.**

Luščenje variant v zapisu

- Identifikacija in štetje primerov, v katerih se izluščena zveza kandidatka pojavlja **tudi** v zapisu skupaj ali z vezajem (zaenkrat niso zajete zveze, ki se v korpusih pojavljajo zgolj v zapisu skupaj in/ali z vezajem).

- Rezultati kažejo, da je variantnost prisotna pri približno četrtini podatkov, zanimivo pa se višja variantnost izkazuje v korpusu Kres.

Zapis	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 %
skupni delež dvojnic oz. trojnic	**29 %**	**25 %**

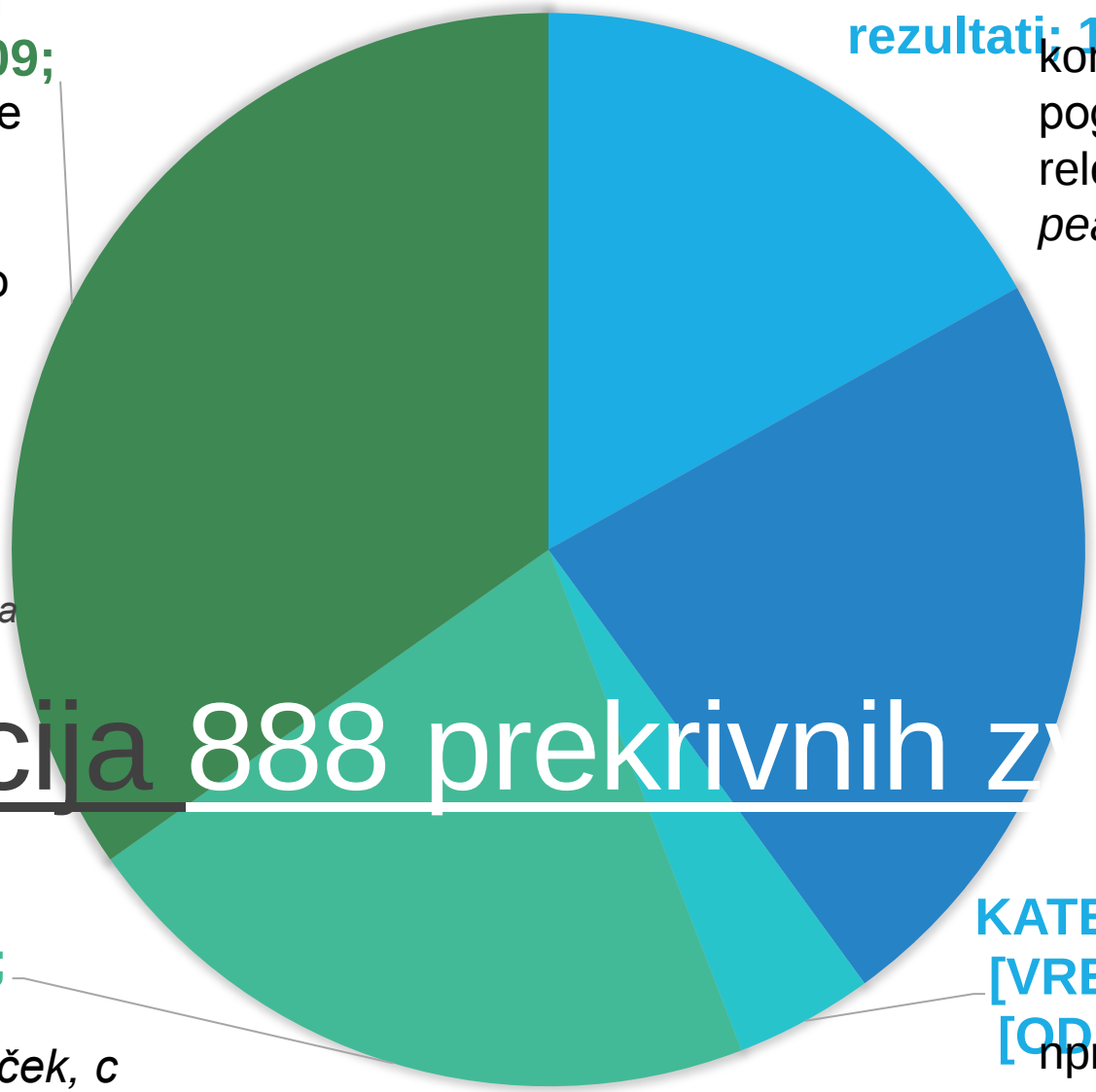
Kategorizacija 888 prekrivnih z

Občna imena z nekratičnim prilastkom; 309; 35%

tako z nesklonljivim samostalniškim prilastkom, ki je bodisi lastno (*android telefon*) bodisi občno ime (*joga studio*), kot tudi zveze z okrajšano prvo sestavino (*info točka*) ali nesklonljivim pridevniškim prilastkom (*mikro podjetje*).

Problem 3: Težave z razmejevanjem skupin!

(Dobrovoljc in Jakop 2011: 114: *eko šola* vs. *mini krilo* vs. *golf igrišče*).



Nerelevantni rezultati; 150; 17%

kombinacije, ki so ustrezale pogojem luščenja, vendar niso relevantne za raziskavo (*york city, pearl jama, družba človek*).

Lastna imena; 205; 23%

(zemljepisna, stvarna), tako domača (*butan plin, ford fiesta*) kot tuja (*financial times*), v podatkih pa se pojavljajo tudi osebna imena (*indiana jones, chuck norris, darth vader, bruce lee, che guevara, james bond, mick jagger, janez janša*)

[IME KATEGORIJE]; [VREDNOST]; [ODSTOTEK]

npr. *after party, bad boy, fair play, press center, team building*

Kratične zveze; 187; 21%

npr. *rtv prispevek, usb ključek, c vitamin, lcd zaslon,* tudi *zf film, fb stran*

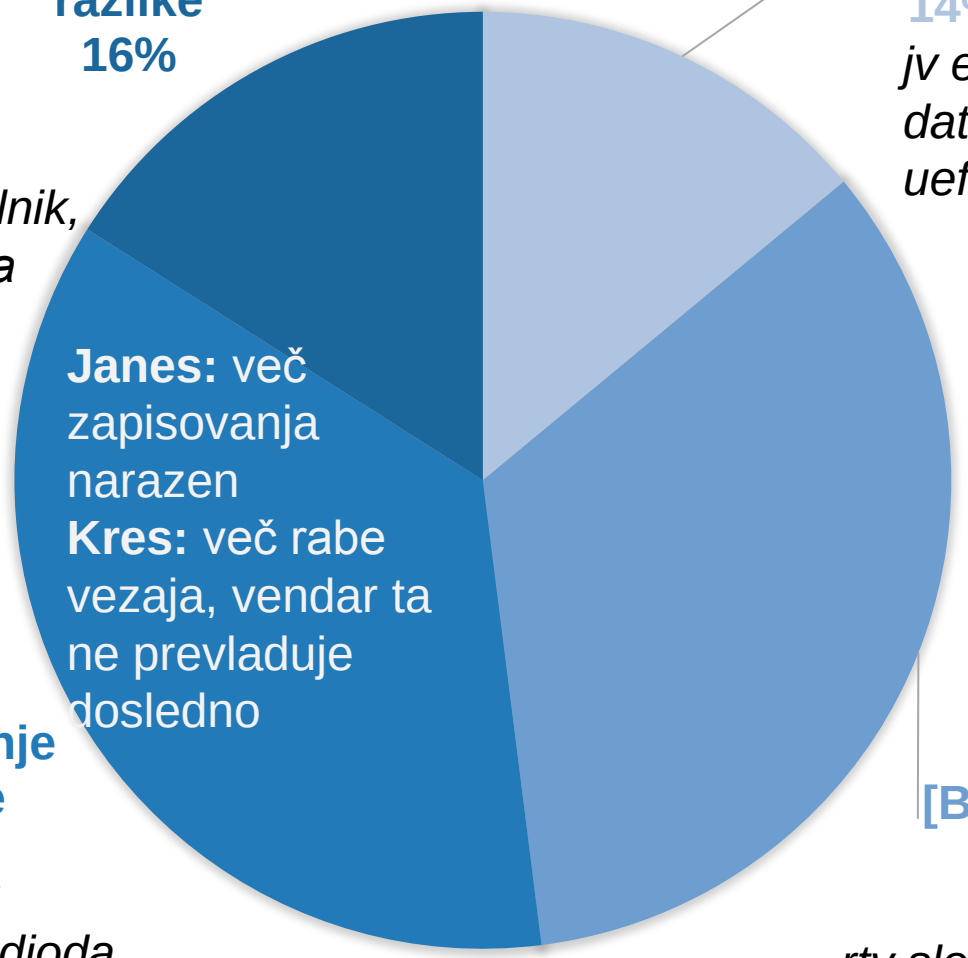
Zapisovanje zvez Janes vs. Kres

- V kolikšni meri se obravnavana korpusa razlikujeta glede trendov v zapisu besednih zvez tipa *USB ključek/USB-ključek* in *joga studio/jogastudio*?

- Za vse ustrezajoče podatke so bila izračunana razmerja, v kolikšnem deležu se posamezna zveza pojavlja zapisana **narazen, skupaj** ali **z vezajem.** Nato smo deleže primerjali med obema korpusoma in zveze razvrstili v štiri skupine:
  a) Ni razlik: zveze, pri katerih **ne prihaja do razlik,** npr. *loto številka, tempera barva, pat pozicija*, ki se v obeh korpusih pišejo izključno narazen.
  b) Majhne razlike: zveze, pri katerih se posamezni deleži **razlikujejo do 25 odstotnih točk,** npr. *pop pevka* se v Janesu zapisuje narazen v 99,3 %, v Kresu pa v 90,2 % primerov.
  c) Srednje razlike: zveze, pri katerih je **razhajanje med 25 in 50 odstotnimi točkami,** npr. *solo petje* se v korpusu Janes zapisuje narazen v 71,7 %, v Kresu v 45,1 % primerov.
  d) Velike razlike: zveze, pri katerih so **razhajanja večja od 50 odstotnih točk,** npr. *lcd zaslon* je v korpusu Janes zapisan narazen v 97,7 %, v Kresu pa v 47 % primerov.

Kratične zveze

Zvez, pri katerih je na prvem mestu kratica, je med podatki 187. Glede na Pravopis (§ 496) naj bi se tovrstni primeri zapisovali z vezajem.



[D] velike razlike 16%

tv oddaja, 3d model, mp3 predvajalnik, iq test, g točka

[A] ni razlik 14%

jv evropa, pdf datoteka, html koda, uefa liga, sv vojna

Janes: več zapisovanja narazen
Kres: več rabe vezaja, vendar ta ne prevladuje dosledno

[C] srednje razlike 36%

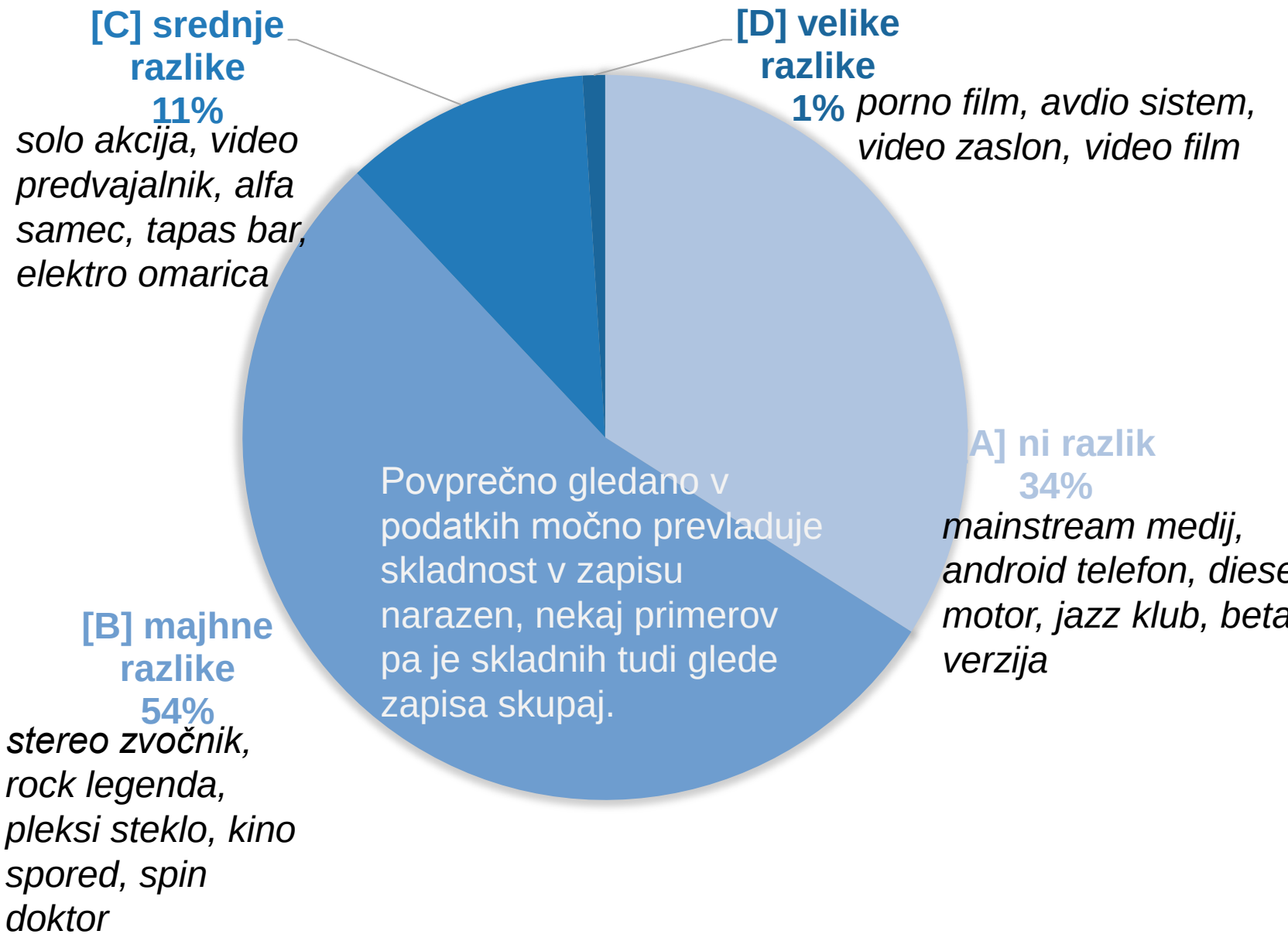
tv program, rtv prispevek, led dioda, usb ključek, c vitamin

[B] majhne razlike 34%

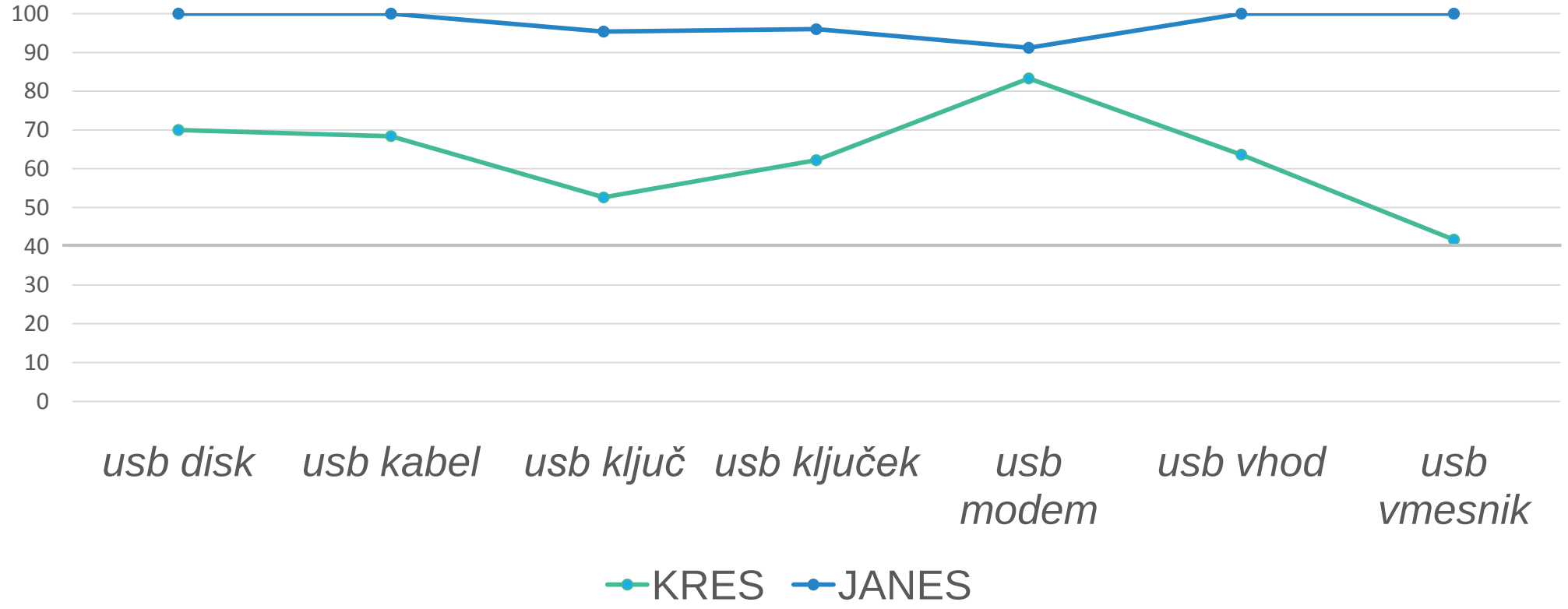
rtv slovenija, eu poslanec, dsg menjalnik, nba liga, sms donacija

Občna imena z nekratičnim prilastkom

Raznovrstnih zvez z nesklonljivim levim prilastkom je med podatki 309. Trenutna jezikovna pravila za te zveze predvidevajo zapis skupaj ali narazen, kar predstavljata (Dobrovoljc in Jakop 2011: 113–122).



Delež zvez s prilastkom *usb*, ki so zapisane narazen

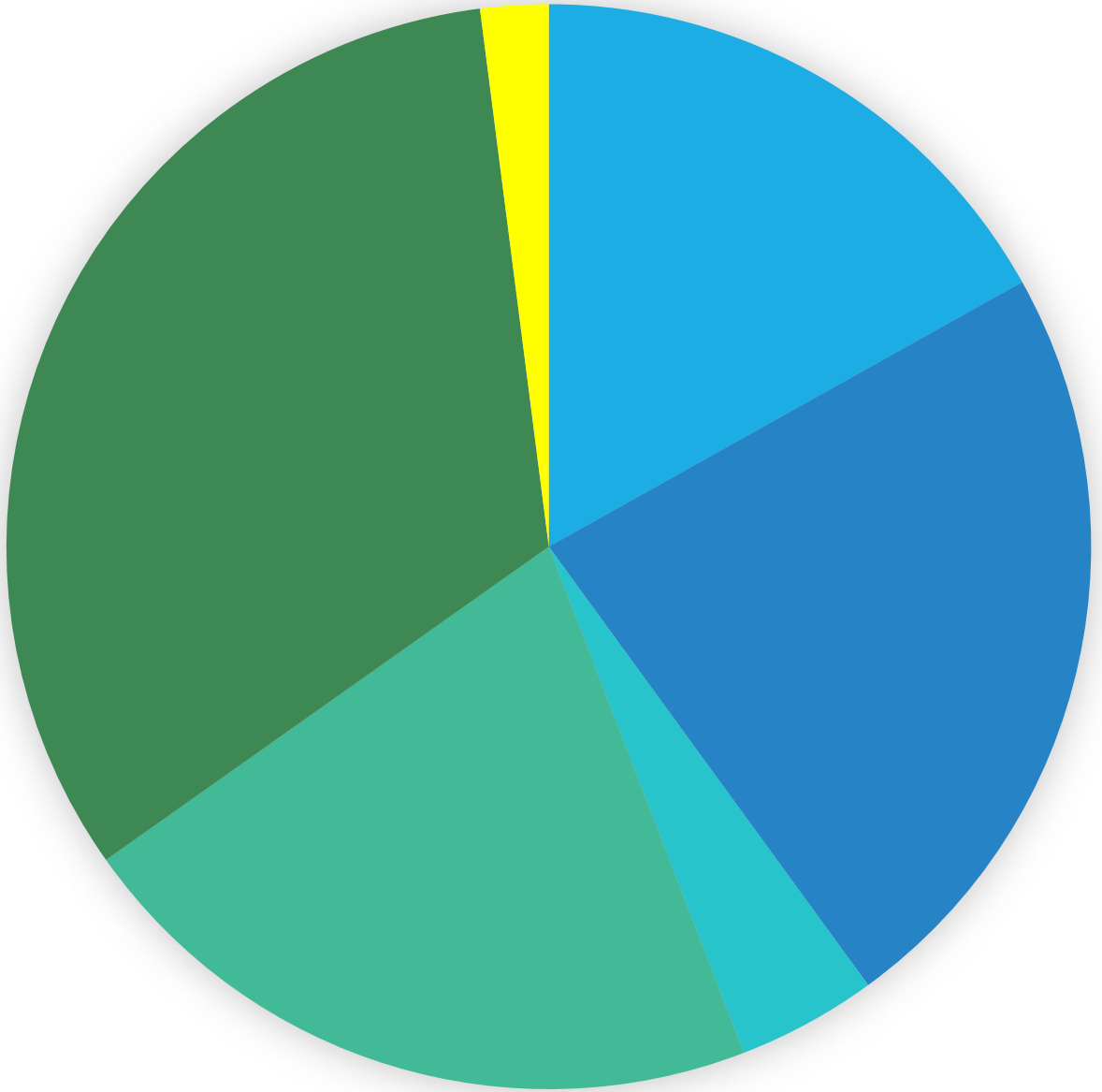


Delež zvez s prilastkom *solo*, ki so zapisane narazen

Kaj pa zapis skupaj?

V Janesu **več kot 50-odstotno pojavitev zapisa skupaj** izkazuje **18 primerov:**

avtocesta, videoposnetek, fotogalerija, videospot, avtošola, avtohiša, kinodvorana, elektromotor, motošport, fotozgodba, videokaseta, turbomotor, betablokator, avtosalon, fotodelavnica, elektroinženir, narkokartel in *videokonferenca*.



Poudarki

1. Glede pojavljanja in rabe obravnavane vrste besed oz. zvez se korpusa Kres in Janes pomembno razlikujeta (opiranje izključno na podatke iz referenčnega korpusa torej lahko vodi do pomanjkljivih zaključkov glede tendenc jezikovne rabe).
2. Označevanje korpusov z različnimi označevalniki otežuje primerjavo podatkov.
3. Nedosleden in nejasen jezikovni opis se prenaša na označevanje korpusnih podatkov in na rezultate luščenja (pri pripravi novega jezikovnega opisa/predpisa, ki temelji na (izluščenih) korpusnih podatkih, je treba to verigo prekiniti).
4. Jezikovna regulacija v smer zapisa skupaj/z vezajem krepi variantnost v jezikovni rabi.
5. Skladenjski vzorec z nesklonljivim levim prilastkom presega primere, na podlagi katerih tipično teče jezikoslovna razprava (primeri tipa *avtocesta, kinodvorana, videokaseta* so v manjšini).
6. Obstoječo jezikoslovno tipologizacijo prilastkov je v praksi težko aplicirati.
7. V splošnem se kaže močan trend k zapisu narazen, ampak variantnost je močno prisotna, tudi na ravni posameznih prilastkov, ki se v različnih zvezah različno zapisujejo.
8. V načrtu imava nadaljnje raziskave (gl. prispevek).

(angora kunec)

Hvala za pozornost!

ŠPELA ARHAR HOLDT IN KAJA DOBROVOLJC

SPELA.ARHAR@TROJINA.SI

KAJA.DOBROVOLJC@TROJINA.SI

Literatura

Helena Dobrovoljc in Nataša Jakop. 2011. *Sodobni pravopisni priročnik med normo in predpisom*. Ljubljana: Založba ZRC.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec in Miro Romih. 2015. Morphological lexicon Sloleks 1.2, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1039.

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, Ralf Steinberger. 2005. Massive multi-lingual corpus compilation: Acquis Communautaire and totale. V *Proceedings of the 2nd Language & Technology Conference*, str. 32–36. Poznan, Poland.

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić, 2015. Gradnja in analiza korpusa spletne slovenščine JANES. V: *Slovnica in slovar - aktualni jezikovni opis (Obdobja 34)*.

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Filozofska fakulteta. V tisku.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik Osme konference Jezikovne tehnologije,* str. 89–94. Ljubljana: Institut Jožef Stefan.

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek in Nanika Holz. 2013. Training corpus ssj500k 1.3, *Slovenian language resource repository CLARIN.SI,* http://hdl.handle.net/11356/1029.

Nataša Logar. 2005. Filter vrečka ali filtervrečka, foto posnetek ali fotoposnetek, ISDN paket ali ISDN-paket? V: M. Jesenšek, ur., *Knjižno in narečno besedoslovje slovenskega jezika*, str. 222–49. Maribor: Slavistično društvo.

Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Nataša Logar. 2012. Razmejitev med besednimi zvezami in zloženkami v sodobnem jezikovnem gradivu. V: N. Jakop in H. Dobrovoljc, ur., *Pravopisna stikanja: Razprave o pravopisnih vprašanjih*, str. 113–23. Ljubljana: Založba ZRC.