cjvt Center za jezikovne vire in tehnologije

KOLOS Kolokacije v slovenščini

arrs JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE

Priprava učne množice za opredelitev kolokativnosti in druge dejavnosti v projektu KOLOS

Jaka Čibej

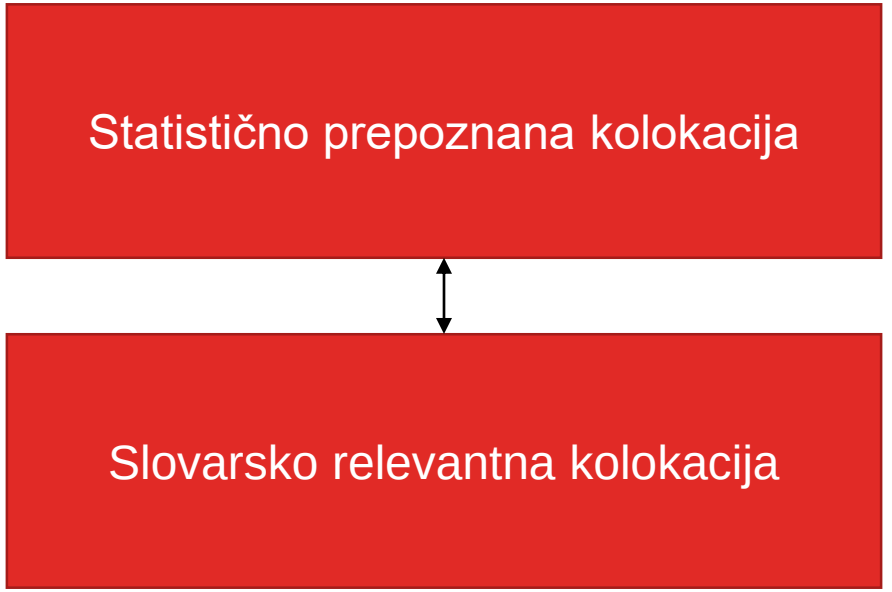
Filozofska fakulteta, Univerza v Ljubljani
Institut "Jožef Stefan"
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Ljubljana, 16. oktober 2018

Kolokativnost in kolokacija

- **Definicije kolokacij**, uporabljene v sorodnih raziskavah in projektih?
- "leksikalno in/ali pragmatično povezane sopojavitve vsaj dveh leksikalnih enot, ki sta med seboj v neposrednem skladenjskem razmerju" [Bartsch 2004]
- "a noticeable arrangement or conjoining of linguistic elements (such as words)" [Merriam Webster]
- "a sequence of words or terms that co-occur more often than would be expected by chance" [Wikipedia]

- **kvantitativno prepoznaven jezikovni pojav**



- **Statistično prepoznane sopojavitve besed?**
  - France Prešeren, 7. festival, večinoma v [Ilirski] Bistrici, mešati hruške [in jabolka]

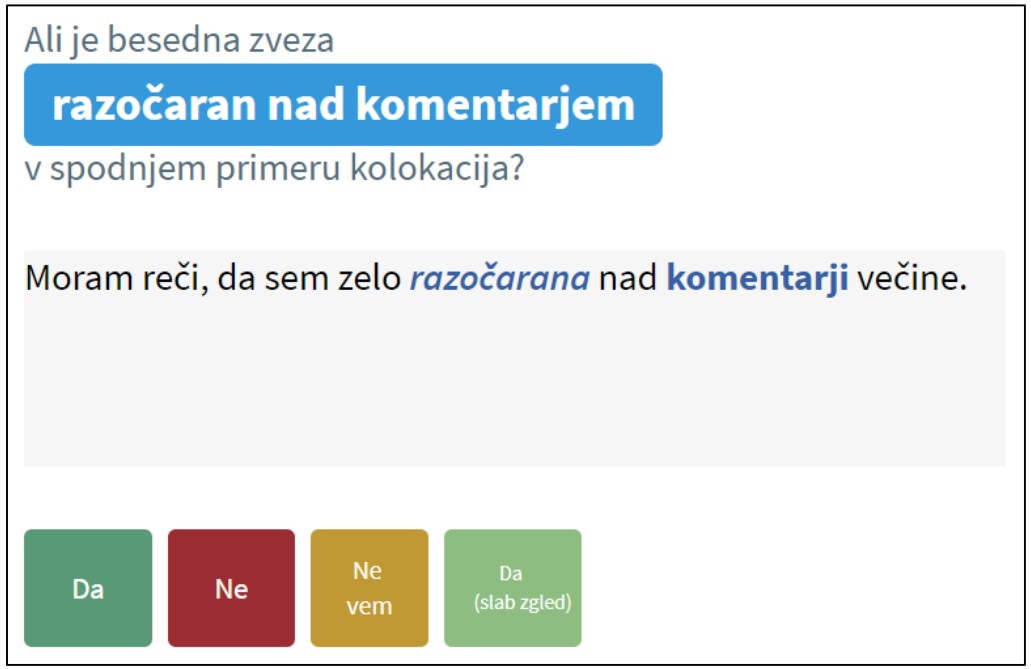
Opredelitev v okviru izgradnje učne množice

- **Okvir gradnje učne množice:**
  - podlaga za razvoj nove metode avtomatskega luščenja kolokacij
  - evalvacija avtomatske metode
- **Dodatna dejavnost:**
  - opredeljevanje kolokativnosti
  - slovarsko relevantna kolokacija?
- Opredelitev kolokativnosti in kolokacij **od spodaj navzgor: od konkretnih primerov k posplošitvam**

- **Izhodiščno vprašanje:** ali med jezikoslovci obstaja dovolj velik konsenz glede tega, kaj je in kaj ni (slovarsko relevantna) kolokacija?

Pilotna množičenjska naloga

- **Ocenjevanje avtomatsko izluščenih kolokacijskih kandidatov**
  - PyBossa, odprtokodna platforma za množičenjske naloge
  - kolokacijski kandidati, izluščeni po različnih vzorcih (npr. pridevnik + samostalnik, glagol + samostalnik v tožilniku)
  - označevalci: 7 jezikoslovcev
  - opcije: Da – Ne – Ne vem – Da (slab zgled)
- **Rezultati**
  - približno 8.800 označenih kandidatov s po 3 odgovori
  - ujemanje:
    - v povprečju 62 % enakih odgovorov med označevalci (med 42 in 76 %)
    - povprečna Cohenova kapa 0,35
  - opazne razlike med različnimi strukturami!



Nadaljevanje označevanja

- **Vzorec 333 iztočnic**
  - izbrane po različnih kriterijih (npr. besedna vrsta, večpomenskost, izvor, pogostost v Gigafidi, število izluščenih kolokacij)
  - čim bolj heterogen nabor → čim bolj reprezentativen vzorec
- **Ocenjevanje kolokacijskih kandidatov izbranih 333 iztočnic**
  - 1 naloga na strukturo (npr. samo *pridevnik* + *samostalnik*)
  - 3 odgovori na mikronalogo, 7 jezikoslovcev
  - več (pod)opcij za odgovore:
    - Da (množina) → *tihotapljena cigareta – tihotapljene cigarete*
    - Da (razširjena kolokacija) → *bitna aplikacija – 16-bitna aplikacija*
    - Ne (struktura) → **polna cvetica* – na travniku, *polnem* pisanih *cvetic*
    - ...

Rezultati – ujemanje

- **Skupaj 17.576 označenih kolokacijskih kandidatov**

- **Ujemanje med označevalci** se razlikuje glede na strukturo.
- Pri nekaterih strukturah je mnogo več razhajanja.

- **Nadaljevanje:**
  - Kje so torej sive cone?
  - Katere strukture so bolj/manj problematične?

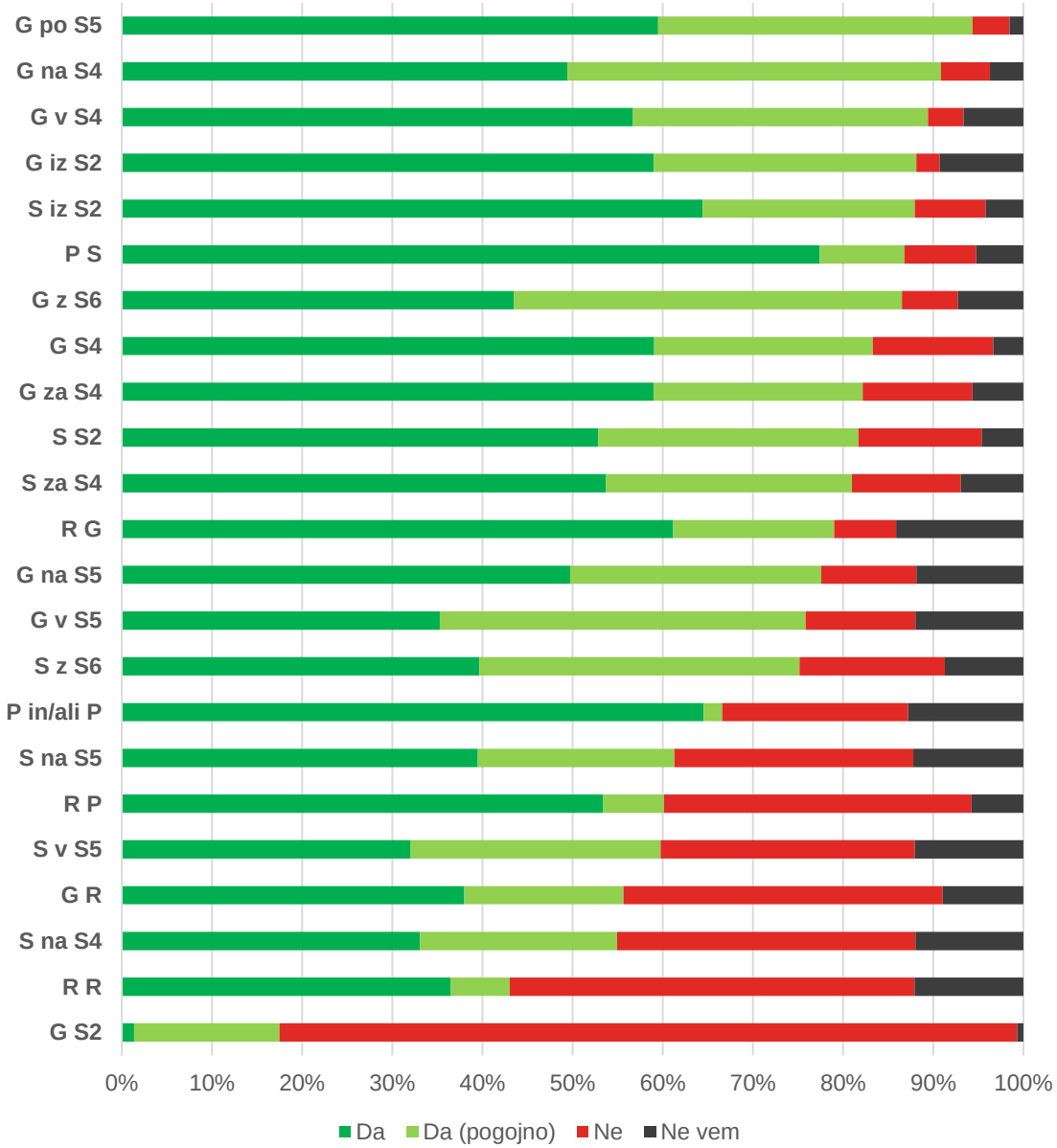
Struktura	Primer	Delež enakih odgovorov	Povprečna Cohenova kapa
G na S4	plezati na jambor	0,8	0,7
G S2	primanjkovati goveda	0,79	0,4
P S	religiozna avtoriteta	0,78	0,42
G po S5	poseči po cigareti	0,73	0,56
G S4	opustiti alkohol	0,73	0,46
…	…	…	…
R R	nadvse burno	0,56	0,4
S v S5	bife v centru	0,5	0,35
S na S5	aplikacija na mobitelu	0,48	0,31
G v S5	delovati v prestolnici	0,46	0,33
P in/ali P	borben in angažiran	0,42	0,08

Rezultati – odgovori

- **Razporeditev odgovorov označevalcev** glede na strukturo kolokacije.

- Strukture z največjim deležem *Da* in *Da (pogojno)*:
  - **G po S5** – *poseči po cigareti*
  - **G na S4** – *plezati na jambor*
  - **G v S4** – *prevesti v francoščino*
  - **P S** – *televizijska cenzura*

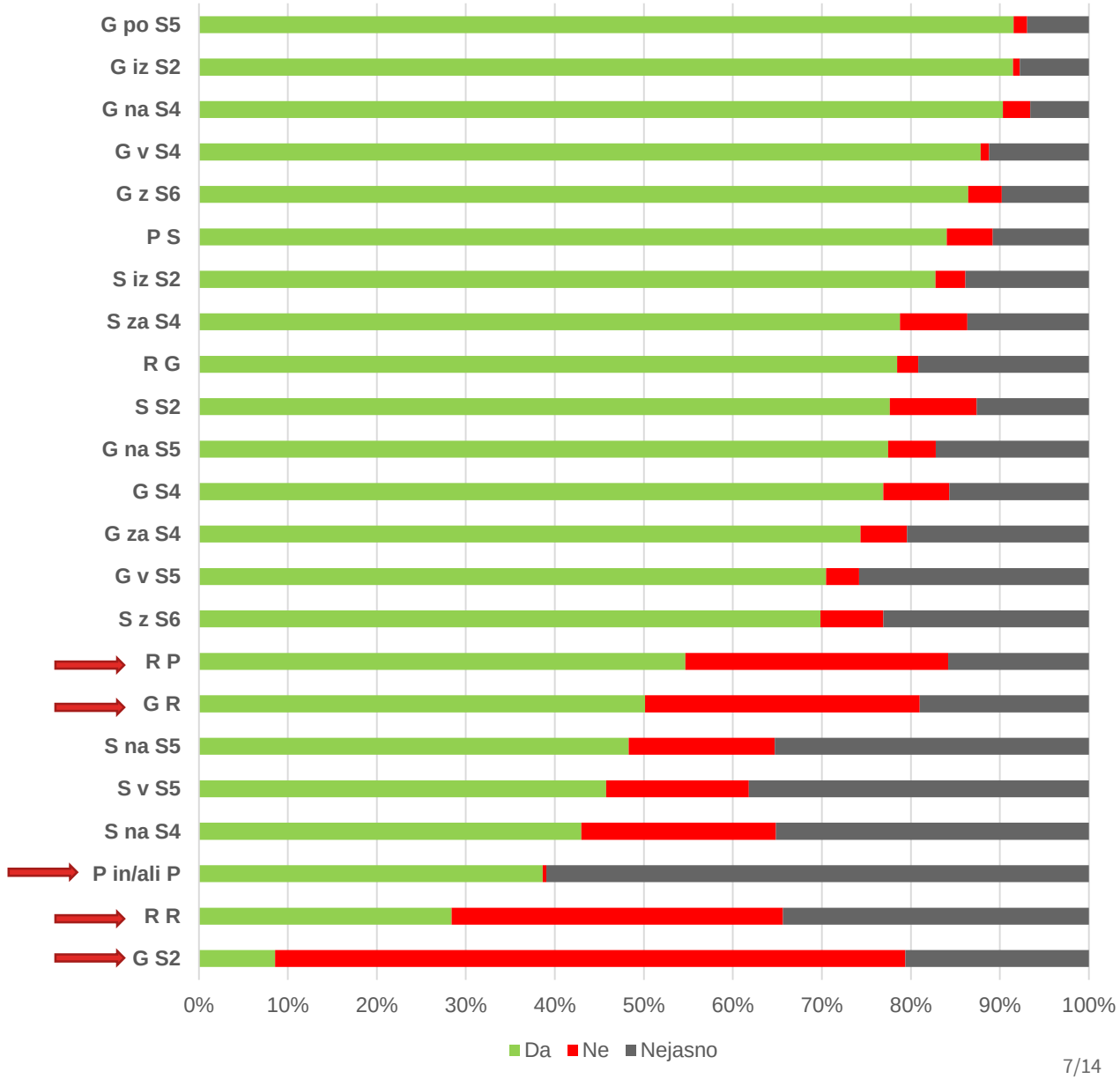
- Strukture z največjim deležem *Ne* in *Ne vem*:
  - **G S2** – *primanjkovati goveda*, **angažirati izvedenca* (tožilnik, ne rodilnik!)
  - **R R** – *dolgo vreti*
  - **G R** – *redko obiskovati*
  - **R P** – *vsestransko angažiran*



Rezultati – odločitve

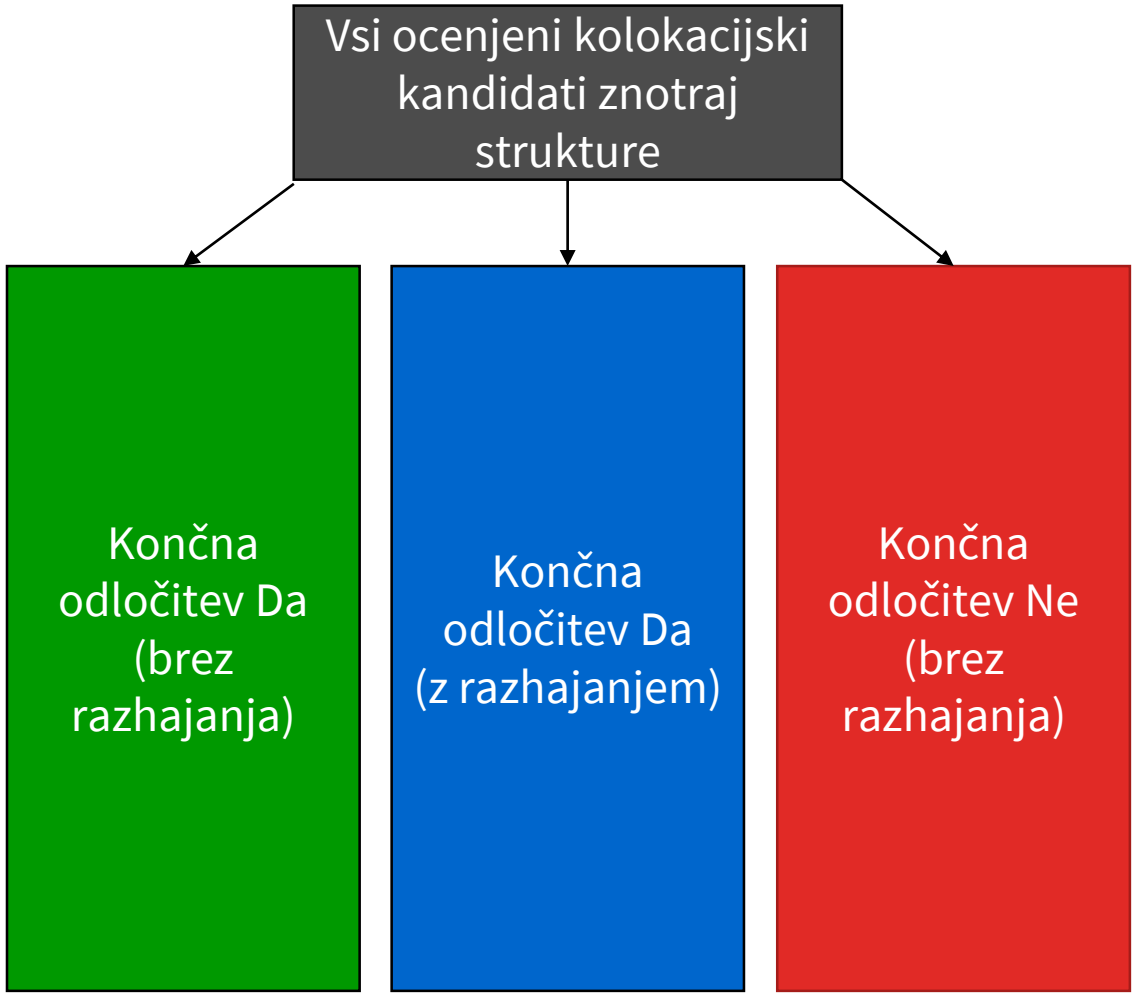
- **Razporeditev končnih odločitev** glede na strukturo kolokacije.
  - večinsko Da = Da
  - večinsko Ne = Ne
  - večinsko Ne vem, mešano = Nejasno

- **Problematične strukture**
  - **P in/ali P** – *prometen in ekološki*
    - *"prometne in ekološke nesreče"*
  - **G S2** – **barvati ograje*
  - **R P** – *dnevno sklenjen [promet]?*
  - **G R** – *boleti enako, prebiti tam?*
  - **R R** – *kdaj zboleti, treba angažirati?*



Pomensko relevantne kolokacije: primer prislovnih struktur

- **Rezultati označevanja kot pomoč pri opredelitvi**
- Analiza vseh kolokacijsko produktivnih struktur s prislovi:
  - **R G** – *treba angažirati*
  - **G R** – *boleti enako*
  - **R R** – *blazno glasno*
  - **P R** – *dostopen brezplačno*
  - **R P** – *pretežno aluminijast*
  - **R in/ali R** – *burno in glasno*
  - **R [predlog] S** – *brezplačno na razpolago*
- Razvrstitev kolokatorjev v **skupine** glede na končno odločitev glede ustreznosti kolokacije
- **Primerjava** – katere vrste prislovov najdemo kot kolokatorje v teh treh skupinah?



Končna odločitev Da (brez razhajanja)

Vrsta prislova	R G	G R	R R	P R	R P	R in/ali R	R [predlog] S
lastnostni	brezplačno, natančno, …	brezplačno, natančno …	strmo, trdno	pokonci	pokončno	vestno, pošteno, pokončno	brezplačno, izjemoma
merni	hudo, blazno, pošteno, močno, znatno	preveč	premalo, dokaj, karseda	/	pošteno, strašno, znatno, hudo, močno	/	nič, precej, četrt, pol
kratnostni	mnogokrat, velikokrat, večkrat	naenkrat, pogosto	/	/	/	/	/
časovni	dnevno, dolgo, kratko, nenehno	dnevno, kratko, občasno	zgodaj, predolgo	/	/	/	/
stopnjevalni	bolj, najbolj	/	/	/	/	/	/
krajevni	doma	doma	/	doma	/	/	doma

Končna odločitev Da (z razhajanjem)

Vrsta prislova	R G	G R	R R	P R	R P	R in/ali R	R [predlog] S
lastnostni	težko	takole	/	dosegljiv, dostopen	/	/	/
merni	tako, večinoma	tako	kar, res, zares, tako, toliko	/	tako, res, večinoma	/	precej, veliko, premalo, preveč
kratnostni/ zaporednostni	enkrat, pogosto, znova, drugič	dvakrat, malokrat	dvakrat, trikrat, kolikokrat	/	/	četrtič	četrtič
časovni	takoj, letos, lani …	takoj, naprej	vedno, danes, doselj, odslej	spomladi	dnevno	spomladi	sinoči, spomladi, dnevno
stopnjevalni	bolj, najbolj *(tip s se/si)*	manj	bolj, najbolj *(tip precej bolj, kar najbolj)*	/	bolj, najbolj *(tip še bolj)*	/	/
kazalni	tu, tukaj, tam	tu, tukaj, tam	tod, tukaj	/	/	/	/
vprašalni	/	/	kje, kam, kako, kdaj, kaj	/	/	/	/

Končna odločitev Da (z razhajanjem)

- **Izhodišče za nadaljnjo razpravo o pomenski relevantnosti skupin prislovov**
  - prislovi kratnosti in pogostnosti – *enkrat*, *pogosto*, *znova*
  - prislovi zaporedja – *prvič*, *drugič*
  - časovni (deikti) – *takoj*, *takrat*, *dnevno*, *letno*
  - kazalni (deikti) – *tukaj*, *tam*
  - merni – *tako*, *toliko*
  - pomensko praznejši merni – *večinoma*, *kar*, *okoli*, *res*
  - vprašalni – *kje*, *kam*, *kaj*, *kako*
  - stopnjevalni – *bolj*, *najbolj*

Končna odločitev Ne (brez razhajanja)

- **Znotrajbesedilna referenčnost**
  - *komentirati kako* – ne bom **komentiral, kako** so pripravljeni
  - *nato barvati* – jih šele **nato barvamo**
- **Nanašalnost naprej**
  - *dati natančno* – se je **dalo natančno** določiti → natančno določiti
  - *zboleti kar* – je **zbolelo kar** šest poštarjev → kar šest
  - *obetati izjemno* – se **obeta izjemno** zanimiv finale → izjemno zanimiv
- **Napačno avtomatsko označevanje strukture:**
  - *odvečno blago* – zavežite konec niti in **odvečno blago** odrežite (samostalnik, ne prislov)
  - *kar gnati* – me je **kar gnalo** naprej (členek, ne prislov)
  - *uspešno doktorski* – v primeru **uspešno** zaključenega **doktorskega** študija
  - *pokončno volanski* – zelo **pokončno** postavljen **volanski** obroč

Sklepi po analizi

- **Slovarsko nerelevantne so kolokacije s prislovi:**
  - v deiktični vlogi
    - *tukaj boleti*, *določiti tam*
  - v vezniški vlogi
    - *prepričati*, *kaj* [je res]
    - *komentirati*, *kako* [so pripravljeni]
  - modalnosti
    - *treba angažirati*
    - *lahko ohladiti*
  - s pomensko oslabljenostjo
    - *večinoma doma*
- **Odločanje na podlagi strukture:**
  - prislovi v vlogi intenzifikatorja
    - *kar prekiniti*
    - ! – *kar pošteno* [načeti]
  - prislovi števniškosti
    - *četrtič doktorirati*
    - ! – *stokrat povedati*

Prihodnje delo

- Analiza vsake kolokacijske strukture in opis kolokativnosti znotraj nje (problemi pri luščenju in pri testiranju drugih metod)
- Izboljšanje luščenja za problematične strukture (npr. G S2, problem prepoznavanja rodilnika)
- Vključevanje novih struktur (npr. osebek + glagol, prvotno izločeno zaradi precejšnjega šuma, npr. *pasti mrak* kot G S4)
- Implementacija ugotovitev v leksikografski delotok in izboljšava podatkov (stopenjskost gesel)
- Uporaba učne množice (17.576 kandidatov) v drugih dejavnostih projekta KOLOS:
  - 2. sklop: uvrščanje kolokatorjev v gruče (ontologija semantičnih tipov, npr. hotel, hiša, dom → zgradba)
  - 3. sklop: primerjava sopomenk s kolokacijami
  - 4. sklop: kolokacijski trendi skozi čas

Hvala za pozornost.

Jaka Čibej
jaka.cibej@ff.uni-lj.si

Center za jezikovne vire in tehnologije

Večna pot 113
1000 Ljubljana
Slovenija

www.cjvt.com
00386 14798299
info@cjvt.si