REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT

Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

Koliko je v slovenskih korpusih besednih vrst, ampak čisto zares?

Simon Krek

Amebis, d.o.o., Kamnik

Institut Jožef Stefan

Edward Sapir: 1921

- Na žalost, ali na srečo, noben jezik ni tiransko konsistenten. Vse slovnice puščajo.
    - Edward Sapir: Language: An Introduction to the Study of Speech. 1921.

Leonard Bloomfield: 1933

- /.../ nemogoče je vzpostaviti povsem konsistentno shemo besednih vrst, ker se razredi besed prekrivajo in križajo.
    - Leonard Bloomfield: Language. 1933.

David Crystal: 1967

- /.../ zaključiti moramo, da so lahko besedne vrste ozke ali široke, kakor narekuje posamezna situacija, in da nobena klasifikacija ni absolutno boljša od druge /.../ različni jezikoslovci bodo za različne namene izdelali bolj ali manj detajlne klasifikacije.
    - David Crystal: English. *Lingua 17*. 1967.

Manning in Schütze: 1999

- Besedna vrsta je pravzaprav kompleksen pojem, kajti motiviran je z različnih izhodišč, npr. s semantičnega (imenovanega tudi pojmovno), distribucijskega skladenjskega ali oblikoslovnega. Takšna pojmovanja besednih vrst so pogosto v konfliktu.
    - Manning in Schütze: Foundations of statistical natural language processing. 1999.

Vrste besed: naloga

- vse besede nekega jezika razdeliti na razrede po izbranih kriterijih

- Kaj je beseda?
- Kaj je razred?
- Kaj je kriterij?

Klasifikacija vs. kategorizacija

- klasifikacija in kategorizacija sta različna koncepta
- klasifikacija je pripisovanje objektov vnaprej definiranim razredom
- kategorizacija je začetna identifikacija teh razredov in se torej mora zgoditi pred klasifikacijo

Gručenje [clustering]

(a) Ena gruča

(b) Dve gruči

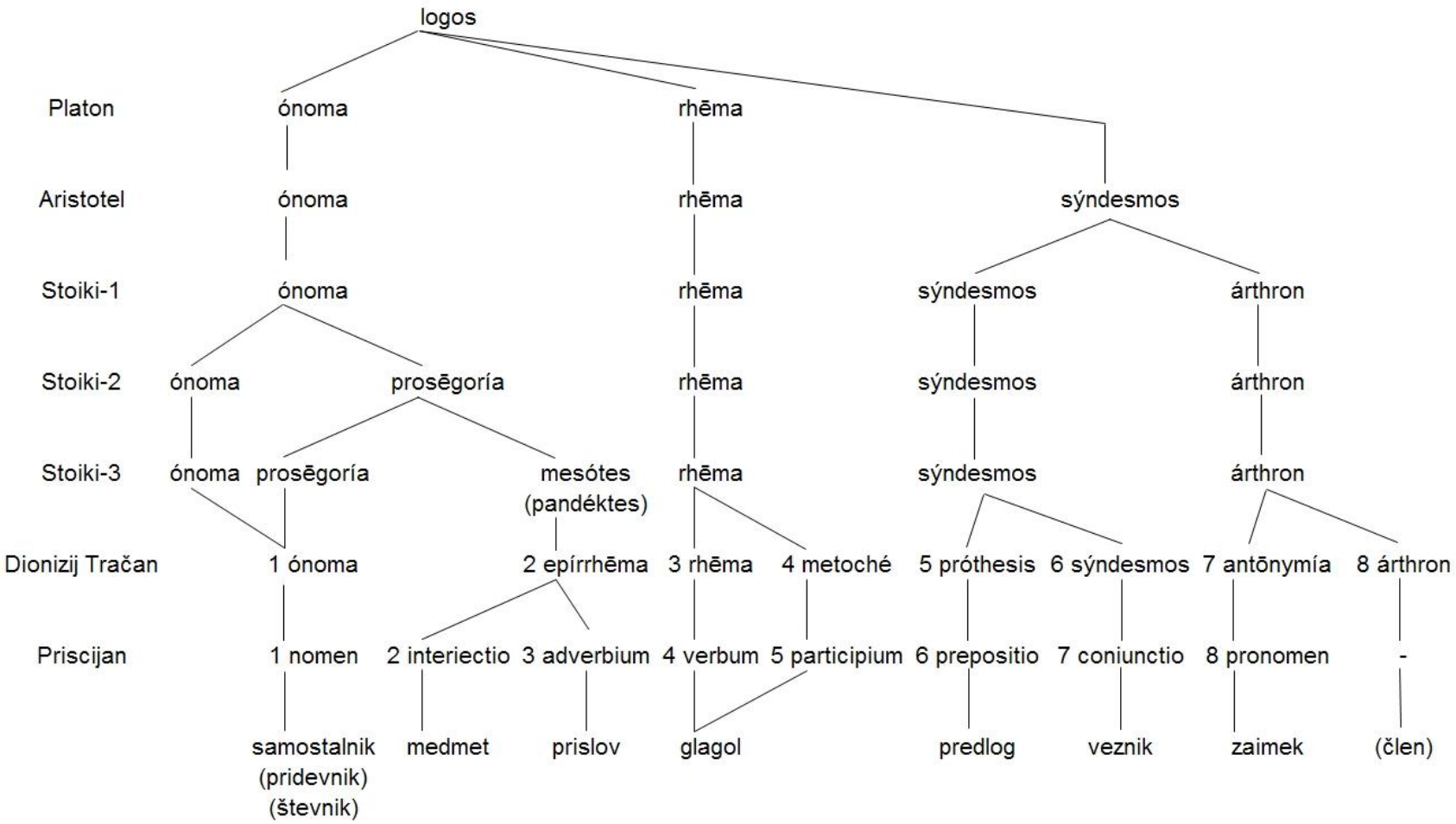
(c) Šriti gruče

(d) Šest gruč

Vreča besed

- "Vreča besed" *[bag-of-words]* je poenostavljajoči privzeti model, uporabljan pri obdelavi naravnih jezikov in informacijskem poizvedovanju *[information retrieval]*. Pri tem modelu je besedilo (npr. stavek ali dokument) predstavljeno kot neurejena zbirka besed, ki ne upošteva niti slovnice niti besednega reda.

Malce zgodovine



Aristotel: 335 pr. n. št.

- Ime (όνομα, ónoma) je sestavljen pomenski izraz brez časovne oznake, čigar sestavni deli nimajo lastnega pomena.
- Glagol (ρήμα, rhêma) je sestavljen pomenski izraz, ki vsebuje časovno oznako in ki v njem – podobno kot v imenu – posamezna sestavina nima lastnega pomena.
- Veznik (σύνδεσμός, syndesmós) je sestavljen izraz, ki nima lastnega pomena in ni primeren, da bi stal sam zase na začetku stavka, /.../

Kriteriji


distribucijski
- skladenjski
predlog
veznik
prislov
zaimek
samostalnik
pridevnik
glagol
števnik
pomenski
oblikovni

Slovenščina I

- (slovnična) kategorizacija & kategorizatorji:
  - Adam Bohorič ...
  - Marko Pohlin ...
  - Jernej Kopitar ...
  - Anton Janežič ...
  - Anton Breznik ...
  - Jože Toporišič ...
- (slovarska) klasifikacija & klasifikatorji:
  - Pleteršnik: Slovensko-nemški slovar ...
  - Slovar slovenskega knjižnega jezika ...
  - Slovenski pravopis ...

Skladenjski premik

	Kopitar (1808/9) Janežič (1854) Breznik (1912) Bajec, Kolarič, Rupel (1956)	Toporišič (2000)
1	samostalnik	samostalniška beseda
2	pridevnik	pridevniška beseda
3	zaimek	povedkovnik
4	številnik	členek
5	glagol	glagol
6	prislov	prislov
7	predlog	predlog
8	veznik	veznik
9	medmet	medmet

Jože Toporišič

- Besedne vrste so v tej knjigi obravnavane kot pojmi za množice besed z enakimi skladenjskimi vlogami in drugimi lastnostmi (npr. tvorjenost, slovnične kategorije, konverznost ipd.). Po tej teoriji je v slovenskem knjižnem jeziku 9 besednih vrst;
    - Jože Toporišič, Slovenska slovnica. 2000.

- /.../ težav in nedoslednosti ter nepopolnosti pa je rešena teorija besednih vrst, ki se opira na skladenjska merila. Taka teorija besedne vrste določa izključno in enotno le po skladenjskih načelih /.../
    - Jože Toporišič, Oblikoslovne razprave. 2003.

Kriteriji


distribucijski
- skladenjski
povedkovnik
členek
pomenski
oblikovni

Protiskladenjski upor

- /.../ povedkovniki oz. povedkovi dopolnilniki niso nič drugega kot pomenske determinante povedkov, zato ostajajo na stavčnočlenski ravni (priložnostna pomenskoskladenjska raba ne more biti besednovrstno odločilna).
    - Andreja Žele, Slovarska obravnava povedkovnika. 2004.

Slovenščina II

- Kategorizacija:
  - Učbeniki:
    - Na pragu besedila, Rokus
    - Z besedo do besede, MK
    - Govorica jezika, Modrijan
    - Besede, DZS
    - ...
  - Učni načrti:
    - Program osnovnošolskega izobraževanja: SLOVENŠČINA (1998)
- Klasifikacija: =Slovenščina I

	besedna vrsta
1	samostalniška beseda
2	pridevniška beseda
3	
4	členek
5	glagol
6	prislov
7	predlog
8	veznik
9	medmet

Marko Stabej: 2010

- Breznikova slovnica nadaljuje slovensko slovnično tradicijo in je hkrati začetek sodobne slovenske slovničarske tradicije 20. in 21. stoletja, ki jo živimo in občutimo še danes. Zanjo je značilno, da je slovnica v temelju delo, namenjeno šolski rabi, natančneje srednješolski populaciji, hkrati pa – bodisi v svoji izvirni obliki bodisi v nadaljnjih metamorfozah – postane tudi osrednje referenčno delo o normi slovenskega knjižnega jezika.
    - Marko Stabej: Slovnica in ideologija. 2010.

Slovenščina III

- (korpusna) kategorizacija & kategorizatorji:
  - Multext-East
  - ISJFR
  - LC-STAR
  - JOS
  - SLON-13
- (korpusna) klasifikacija & klasifikatorji:
  - označevalnik TnT
  - označevalnik TreeTagger
  - označevalnik Amebis
  - označevalnik SLON-13
  - označevalnik SSJ

št	kategorija	JOS	Multext-East	LC-STAR	ISJFR	SLON-13
1	samostalnik	+	+	+	+	+
	lastno ime	Ø	Ø	Ø	+	Ø
2	glagol	+	+	+	+	+
	pomožni glagol	Ø	Ø	+	Ø	+
3	pridevnik	+	+	+	+	+
4	zaimek	+	+	+	+	+
5	števnik	+	+	+	+	+
6	prislov	+	+	+	+	+
7	veznik	+	+	+	+	+
8	predlog	+	+	+	+	+
9	medmet	+	+	+	+	+
10	členek	+	+	+	+	+
11	okrajšava	+	+	+	+	+
	neuvrščeno	+	+	Ø	Ø	+

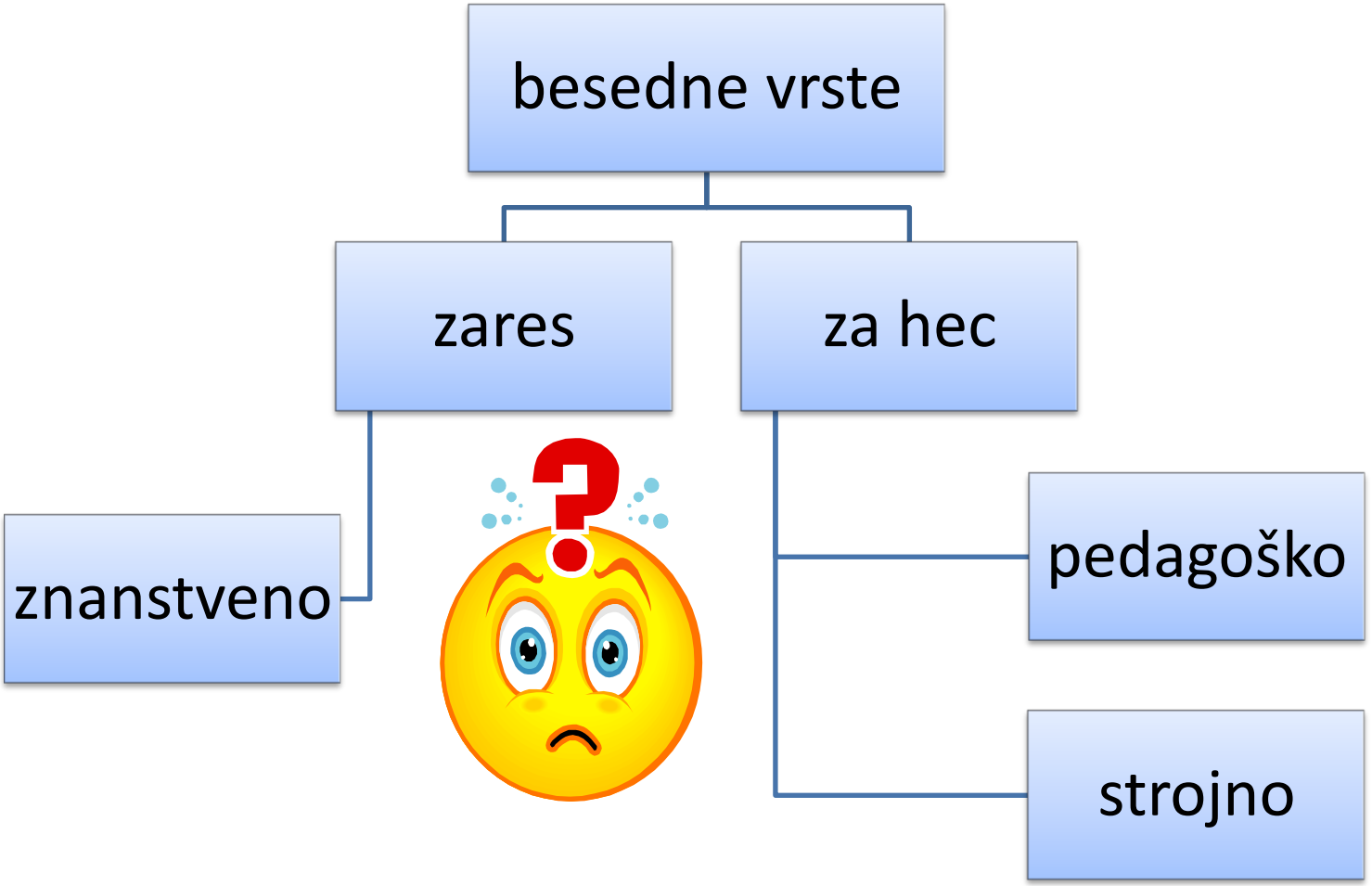
"Fleksemsko" izhodišče

		liczba	przypadek	rodzaj	osoba	stopień	aspekt	zaneg.	akcent.	poprzyim.	akomod.	aglutyn.	wokal.
		število	sklon	spol	oseba	stopnja	vid	nikalnost	klitičnost	naveznost (predlog)	ujemalnost	naveznost (biti)	vokalizacija
subst	samostalnik	⊕	⊕	⊙									
depr	sam. (slabšalno)	⊙	⊕	⊙									
num	glavni številnik	⊙	⊕	⊕							⊕		
numcol	≈ ločilni številnik	⊙	⊕	⊙							⊕		
adj	pridevnik	⊕	⊕	⊕		⊕							
adja	≈ obpridevniški prid.												
adjp	≈ popredložni prid.												
adv	prislov					⊕							
ppron12	zaimek netretjeosebni	⊙	⊕	⊕	⊙				⊕				
ppron3	zaimek tretjeosebni	⊕	⊕	⊕	⊙				⊕	⊕			
siebie	zaimek sebe		⊕										
fin	ne-pretekla oblika	⊕			⊕		⊙						
bedzie	prihodnja oblika "biti"	⊕			⊕		⊙						
aglt	navezni "biti"	⊕			⊕		⊙						⊕
praet	deležnik na -l	⊕		⊕			⊙					⊕	
impt	velelnik	⊕			⊕		⊙						
imps	brezosebna oblika						⊙						
inf	nedoločnik						⊙						
pcon	≈ deležje (sed.)						⊙						
pant	≈ deležje (pret.)						⊙						
ger	glagolnik	⊕	⊕	⊙			⊙	⊕					
pact	≈ tvorni deležnik	⊕	⊕	⊕			⊙	⊕					
ppas	≈ trpni deležnik	⊕	⊕	⊕			⊙	⊕					
winien	"winienski" element	⊕		⊕			⊙						
pred	povedkovnik												
prep	predlog		⊙										
conj	veznik												
qub	členek-prislov												
xxs	tujej. samostalniki	⊕	⊕	⊙									
xxx	tujejezični drugi												
ign	neznano												
interp	ločilo												

št	oznaka	poljsko	slovensko
1	subst	rzeczownik	samostalnik
2	depr	rzeczownik deprecjatywny	samostalnik (slabšalno)
3	num	liczebnik główny	glavni številnik
4	numcol	liczebnik zbiorowy	≈ ločilni številnik
5	adj	przymiotnik	pridevnik
6	adja	przymiotnik przyprzym.	≈ obpridevniški pridevnik
7	adjp	przymiotnik poprzyim.	≈ popredložni pridevnik
8	adv	przysłówek	prislov
9	ppron12	zaimek nietrzecioosobowy	zaimek netretjeosebni
10	ppron3	zaimek trzecioosobowy	zaimek tretjeosebni
11	siebie	zaimek siebie	zaimek sebe
12	fin	forma nieprzeszła	ne-pretekla oblika
13	bedzie	forma przyszła być	prihodnja oblika "biti"
14	aglt	aglutynant być	navezni "biti"
15	praet	pseudoimiesłów	deležnik na -l
16	impt	rozkaźnik	velelnik

št	oznaka	poljsko	slovensko
17	imps	bezosobnik	brezosebna oblika
18	inf	bezokolicznik	nedoločnik
19	pcon	im. przys. współczesny	≈ deležje, ki izraža sedanjost [deležje na -č]
20	pant	im. przys. uprzedni	≈ deležje, ki izraža preteklost [deležje na -ši]
21	ger	odsłownik	glagolnik
22	pact	im. przym. czynny	≈ tvorni deležnik [deležnik na -č]
23	ppas	im. przym. bierny	≈ trpni deležnik [deležnik na –n/-t]
24	winien	winien	"winienski" element
25	pred	predykatyw	povedkovnik
26	prep	przyimek	predlog
27	conj	spójnik	veznik
28	qub	kublik	členek-prislov
29	xxs	ciało obce nominalne	tujejezični samostalniki
30	xxx	ciało obce luźne	tujejezični drugi
31	ign	forma nierozpoznana	neznano
32	interp	interpunkcja	ločilo

Slovenščina 2011

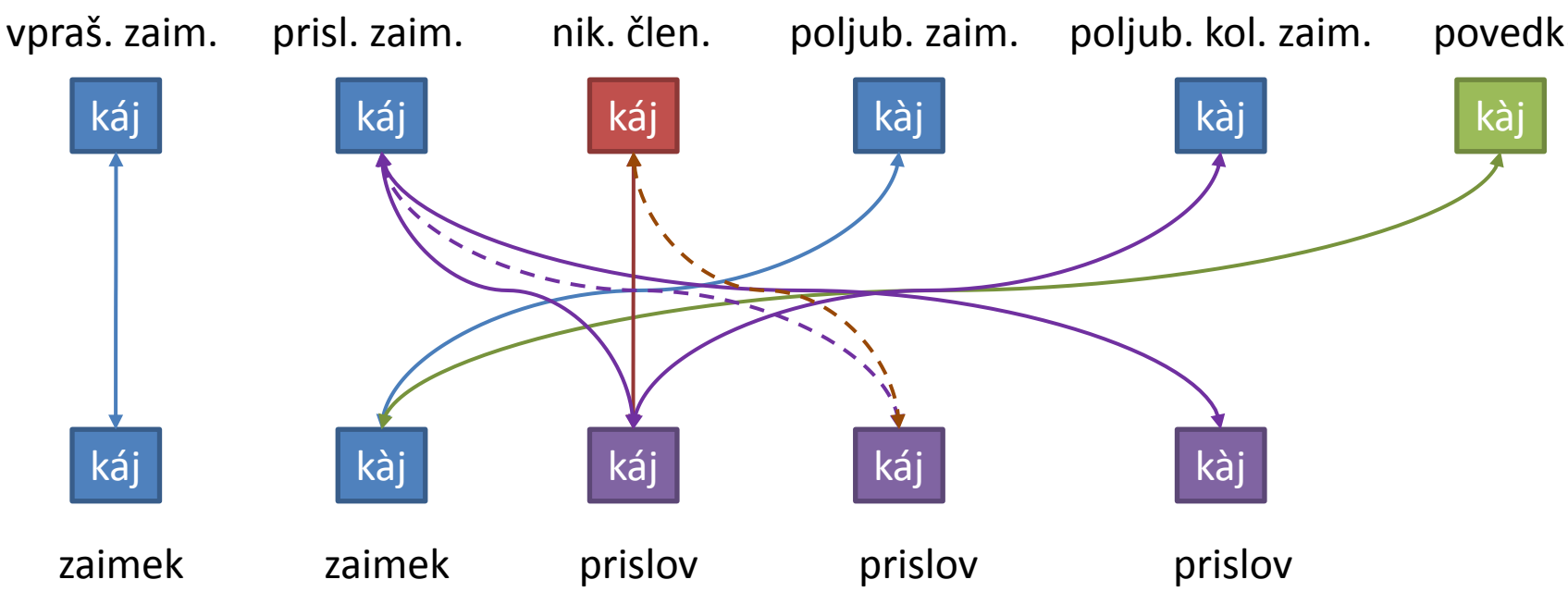


Klasifikacija

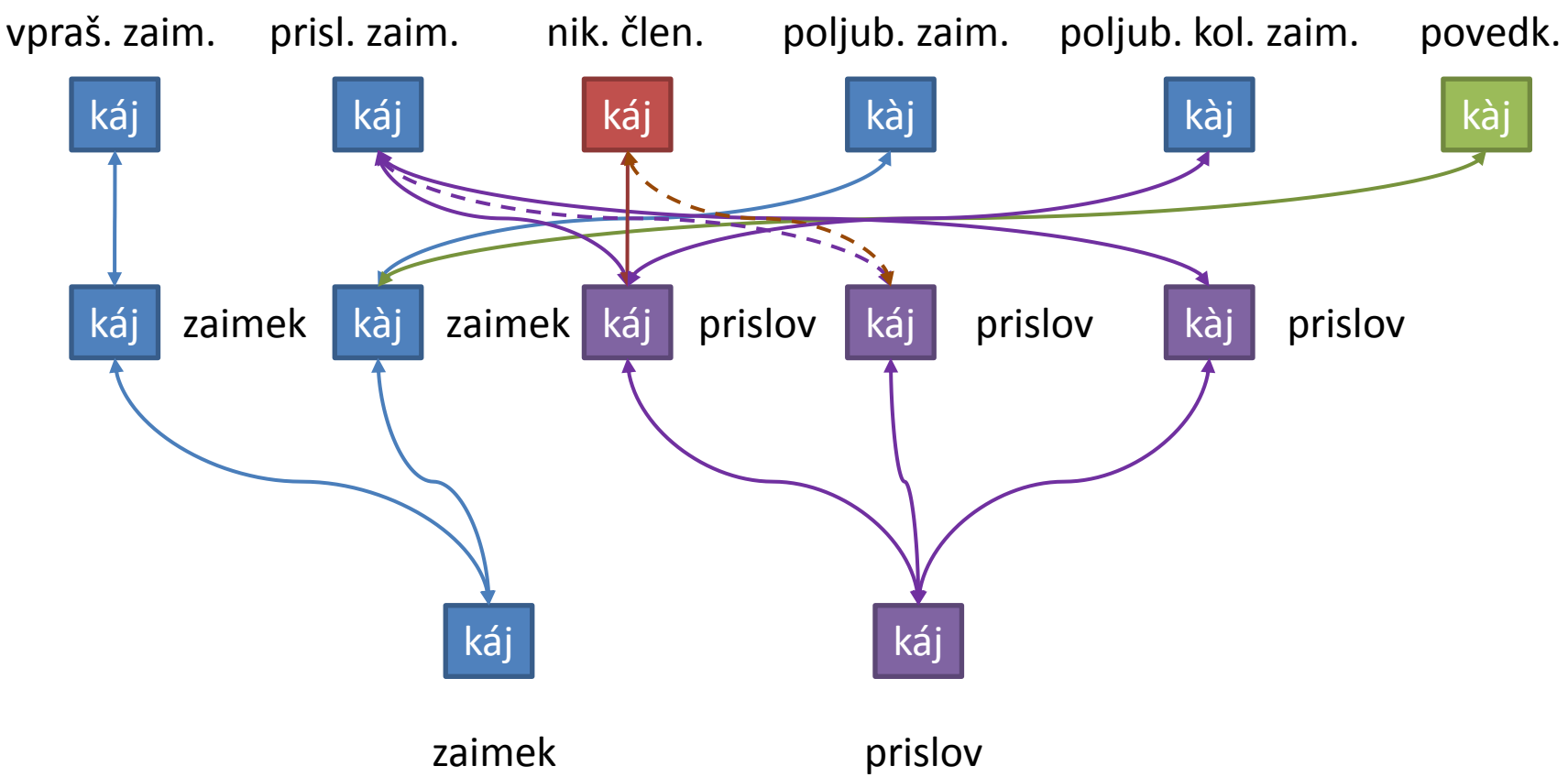
- klasifikacija je pripisovanje objektov vnaprej definiranim razredom

- znanstveno, kanonično: SSKJ, SP
- pedagoško: SSKJ, SP
- strojno: FidaPLUS, Nova beseda itd.

SP – SSKJ


vpraš. zaim.
prisl. zaim.
nik. člen.
poljub. zaim.
poljub. kol. zaim.
povedk.
káj
káj
káj
kàj
kàj
kàj
káj
kàj
káj
káj
kàj
zaimek
zaimek
prislov
prislov
prislov

SP – SSKJ – JOS


vpraš. zaim.
prisl. zaim.
nik. člen.
poljub. zaim.
poljub. kol. zaim.
povedk.
káj
káj
káj
kàj
kàj
kàj
káj
zaimek
kàj
zaimek
káj
prislov
káj
prislov
kàj
prislov
káj
káj
zaimek
prislov

Koliko je … ?

Nekje med tri ...

jos100k	pojavnice	leme	št. p. na lemo
samostalnik	28.865	8.184	3,5
glagol	18.712	2.114	8,8
pridevnik	11.405	3.362	3,4
predlog	10.612	49	216,6
veznik	8.979	42	213,8
zaimek	7.709	76	101,4
prislov	6.072	835	7,3
členek	3.413	42	81,3
števnik	3.020	772	3,9
nedoločeno	732		
okrajšava	455		
medmet	29		
skupaj	100.003	16.692	71,1

in dvaintrideset!

		liczba	przypadek	rodzaj	osoba	stopień	aspekt	zaneg.	akcent.	poprzyim.	akomod.	aglutyn.	wokal.
		število	sklon	spol	oseba	stopnja	vid	nikalnost	klitičnost	naveznost (predlog)	ujemalnost	naveznost (biti)	vokalizacija
subst	samostalnik	⊕	⊕	⊙									
depr	sam. (slabšalno)	⊙	⊕	⊙									
num	glavni številnik	⊙	⊕	⊕							⊕		
numcol	≈ ločilni številnik	⊙	⊕	⊙							⊕		
adj	pridevnik	⊕	⊕	⊕		⊕							
adja	≈ obpridevniški prid.												
adjp	≈ popredložni prid.												
adv	prislov					⊕							
ppron12	zaimek netretjeosebni	⊙	⊕	⊕	⊙				⊕				
ppron3	zaimek tretjeosebni	⊕	⊕	⊕	⊙				⊕	⊕			
siebie	zaimek sebe		⊕										
fin	ne-pretekla oblika	⊕			⊕		⊙						
bedzie	prihodnja oblika "biti"	⊕			⊕		⊙						
aglt	navezni "biti"	⊕			⊕		⊙						⊕
praet	deležnik na -l	⊕		⊕			⊙					⊕	
impt	velelnik	⊕			⊕		⊙						
imps	brezosebna oblika						⊙						
inf	nedoločnik						⊙						
pcon	≈ deležje (sed.)						⊙						
pant	≈ deležje (pret.)						⊙						
ger	glagolnik	⊕	⊕	⊙			⊙	⊕					
pact	≈ tvorni deležnik	⊕	⊕	⊕			⊙	⊕					
ppas	≈ trpni deležnik	⊕	⊕	⊕			⊙	⊕					
winien	"winienski" element	⊕		⊕			⊙						
pred	povedkovnik												
prep	predlog		⊙										
conj	veznik												
qub	členek-prislov												
xxs	tujej. samostalniki	⊕	⊕	⊙									
xxx	tujejezični drugi												
ign	neznano												
interp	ločilo												

Zaključki

- nabor besednih vrst in njihovih lastnosti v korpusih ima praktično naravo – za usodnejše jezikoslovne trditve o jeziku je treba raziskati izbrano jezikoslovno témo takó, da zgolj vzamemo v obzir izbrani sistem kategorizacije
- korpusni nabori oznak to usodo pravzaprav delijo z vsemi oblikoslovnimi opisi – noben ni edini pravi, gre za premikanje fokusa z uporabo različnih kriterijev pri različnih kategorizacijah za različno rabo: npr. pri poučevanju, korpusnem označevanju, analizi diskurza itd. itd.

Alternativi

- prakse so različne: možno je najti konsenz glede osnovnih besednovrstnih kategorij, ostale razrede pa uvrstiti na en nivo nižje ter uskladiti pedagoško prakso s to shemo zaradi lažjega razumevanja podatkov
- alternativa je, da se modela še bolj razdružita, s čimer se zmanjša možnost, da bo laični uporabnik napačno razumel razlike med šolsko informacijo in splošnimi jezikovnimi viri