Priprema *usporedivih* korpusa za *usporedbu*

Ivana Lalli Paćelat
Sveučilište Jurja Dobrile u Puli

Uvod

- važnost i sustavno planiranje izrade skupa oznaka u skladu sa standardima

- preduvjet usporedivosti među korpusima

na unutarjezičnoj i na međujezičnoj razini

- Što kada se takvi skupovi oznaka ne koriste?

Priprema korpusa za istraživanje:

Analiza zakonodavnopravnoga stila hrvatskog i talijanskog jezika: unutarjezična, međujezična i prijevodna perspektiva

Za analizu registra neophodno je imati (Biber 1995):

1) komparativni pristup,
2) kvantitativnu analizu (*distribucija vrsta riječi: omjer imenica i glagola, distribucija glag. oblika, distribucija kateg. lica, distribucija modalnih gl., distribucija vrsta zamjenica, prijedloga, veznika i dr., omjeri pojavnica i različnica, leksičke gustoće, leksičkoga bogatstva;)*
3) reprezentativni uzorak.

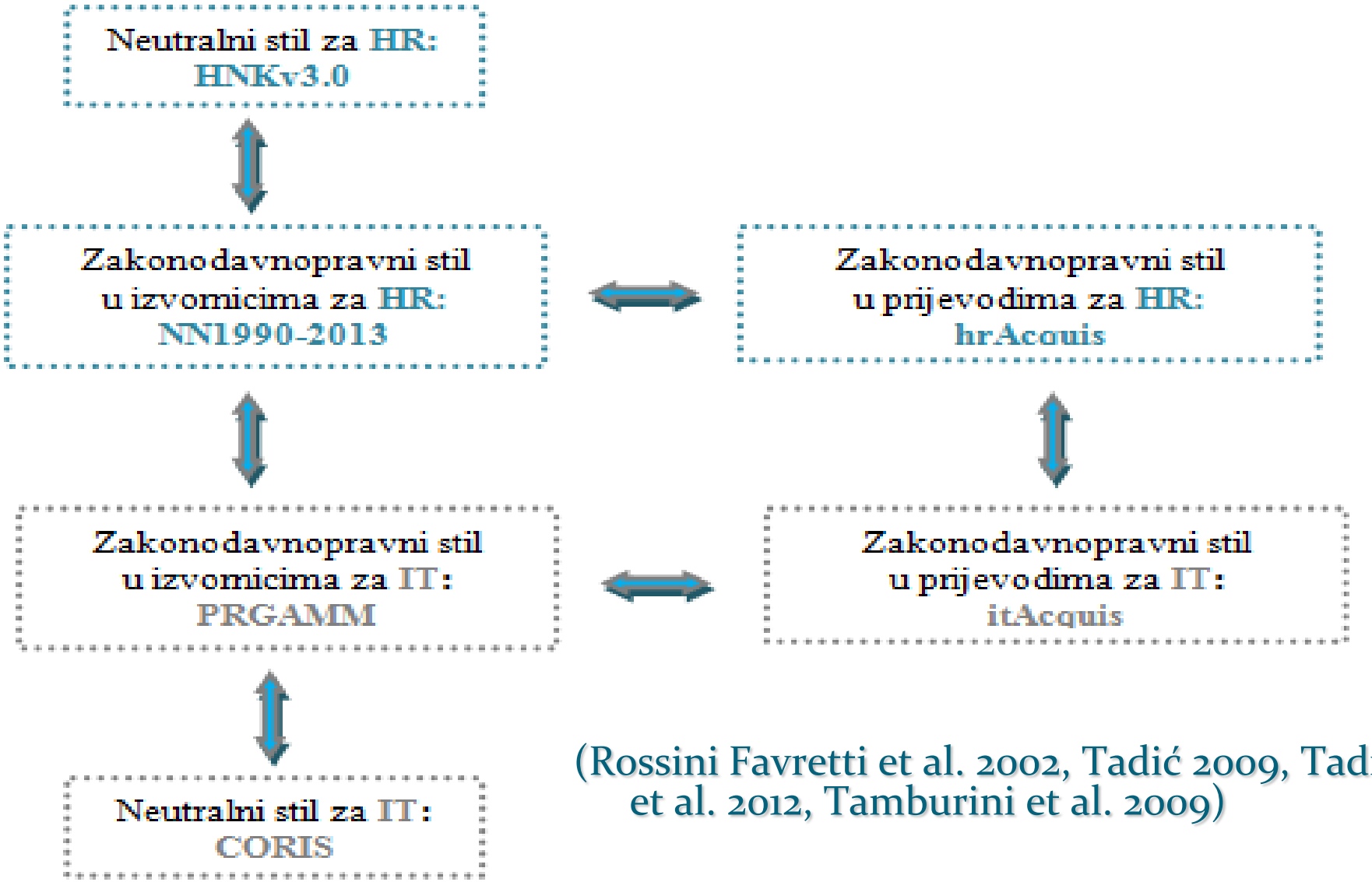
Korpusi

- šest različitih korpusa odnosno četiri vrste korpusa:

1. referentni korpusi: HNK v3.0 i CORIS,
2. dvojezični usporedivi korpus: potkorpus HNK v3.0, NN1990-2013 i potkorpus CORIS-a, PRGAMM,
3. jednojezični usporedivi korpusi: NN1990-2013 i hrAcquis; PRGAMM i itAcquis,
4. usporedni korpus: hrAcquis i itAcquis.*

(Rossini Favretti et al. 2002, Tadić 2009, Tadić et al. 2012, Tamburini et al. 2009)

Komparativni pristup



(Rossini Favretti et al. 2002, Tadić 2009, Tadić et al. 2012, Tamburini et al. 2009)

Provjera usporedivosti među korpusima

- **Korpusi u sličnim istraživanjima !!!**

(Cortelazzo 2013, Neumann, 2013; Teich, 2003; Venturi 2011, Xiao 2010 i dr.)

- **Provjera usporedivosti među korpusima**

- HNK & CORIS / NN1990–2013 & PRGAMM / hrAcquis & itAcquis
  - raspon, opseg, struktura
  - označavanje ?
    - POS/ MSD označavanje

Analiza skupova oznaka

- Skupovi oznaka:

  1) HNKv3.0, NN1990–2013 i hrAcquis

  MULTEXT*-East (v4.0) specifikacija (Tadić 1998, Erjavec 2004)

  2) CORIS i PRGAMM

  !!! **EAGLES-like tagset for Italian** (Tamburini 2000, Monachini 1995)

  3) itAcquis → **Freeling tagset** (Prokopidis i suradnici 2011)

- Razlika u razini označavanja!!!

**MULTEXT preporuka razrađena u suradnji s EAGLES inicijativom iz 1996.*

Potreba za usklađivanjem oznaka

CORIS

Oznaka	Opis
V_AVERE	svi oblici (pomoćnoga) glagola *avere* (hr.*imati*)
V_ESSERE	svi oblici (pomoćnoga) glagola *essere* (hr. *biti*)
V_MOD	svi oblici glagola *potere, dovere, volere* (hr. *moći, morati, htjeti*)
V_PP	glagolski pridjevi[317]
V_GVRB	općeniti[318] glagolski oblici
V_CLIT	združeni[319] glagolski oblici (npr. *andarci, dimmelo*; hr. *ići tamo, reci mi to*)

itAcquis

C	*Vrijednost*	*C*	*Vrijednost*	*C*	*Vrijednost*	*C*	*Vrijednost*
1	**Vrsta**	**2**	**Oblik**	**3**	**Vrijeme**	**4**	**Lice**
V	glavni glagol	**i**	indikativ	**p**	prezent	**1**	1.lice
VA	pomoćni glagol	**m**	imperativ	**i**	imperfekt	**2**	2.lice
VM	modalni glagol	**d**	kondicional	**s**	perfekt	**3**	3.lice
		c	konjuktiv	**f**	futur		
		f	infinitiv				
		g	glagolski prilog (*gerundio*)				
		p	glagolski pridjev (*participio*)				

HNKv3.0, NN1990-2013, hrAcquis

C	*Vrijednost*	*C*	*Vrijednost*	*C*	*Vrijednost*	*C*	*Vrijed*	*C*	*Vrijed.*	*C*	*Vrijed.*	*C*	*Vrijed*	*C*	*Vrijed.*
1	**Vrsta**	**2**	**Oblik**	**3**	**Vrijeme**	**4**	**Lice**	**5**	**Broj**	**6**	**Rod**	**7**	**Stanje**	**8**	**Niječni oblik**
m	glavni glagol	**i**	indikativ	**p**	prezent	**1**	1.lice	**s**	jednina	**m**	muški	**a**	radno	**n**	ne
a	pomoćni glagol	**m**	imperativ	**i**	imperfekt	**2**	2.lice	**p**	množina	**f**	ženski	**p**	trpno	**y**	da
o	modalni glagol	**c**	kondicional	**f**	futur	**3**	3.lice			**n**	srednji				
c	kopula	**n**	infinitiv	**s**	perfekt										
		p	glagolski pridjev i prilog	**l**	pluskvam-perfekt										
				a	aorist										

Usklađivanje oznaka I

- svođenje na (samo) zajedničke POS/MSD oznake
  - odabir : MULTEXT-East (v4.0) specifikacija (Tadić 1998, Erjavec 2004)

složenije pretrage s RI!!

IT-ACQUIS	CORIS	HNK	Odabrana oznaka za ovaj rad i objašnjenje	
V	V	V	V	glagol
VA	V_ESSERE	Va	Va	pomoćni glagol
	V_AVERE	Vc		
VM	V_MOD	Vm	Vm	glagolski oblik suznačnoga ili samoznačnoga glagola
V	V_GVRB			
	V_CLIT			
Vp	V_PP			

Tablica 1: Prijedlog zajedničkih oznaka za glagole

Usklađivanje oznaka II

Razlike u poimanju i postojanju gramatičkih kategorija !

itACQUIS			CORIS	HNK	Odabrana oznaka za ovaj rad i objašnjenje	
S			N	N	N	imenica
Ss	Sp	Sn	NN	Nc	Nc	opća imenica
SP			NN_P	Np	Np	vlastita imenica
SWs	SWp	SWn				
B			ADV	R	R	prilog
				Qr		
BN				Qz		

Tablica 2: Prijedlog zajedničkih oznaka za imenice i priloge

- status čestica !!!
- rješenje: uključivanje dijela pojavnica s oznakom čestica koje se u talijanskome jeziku smatraju prilozima

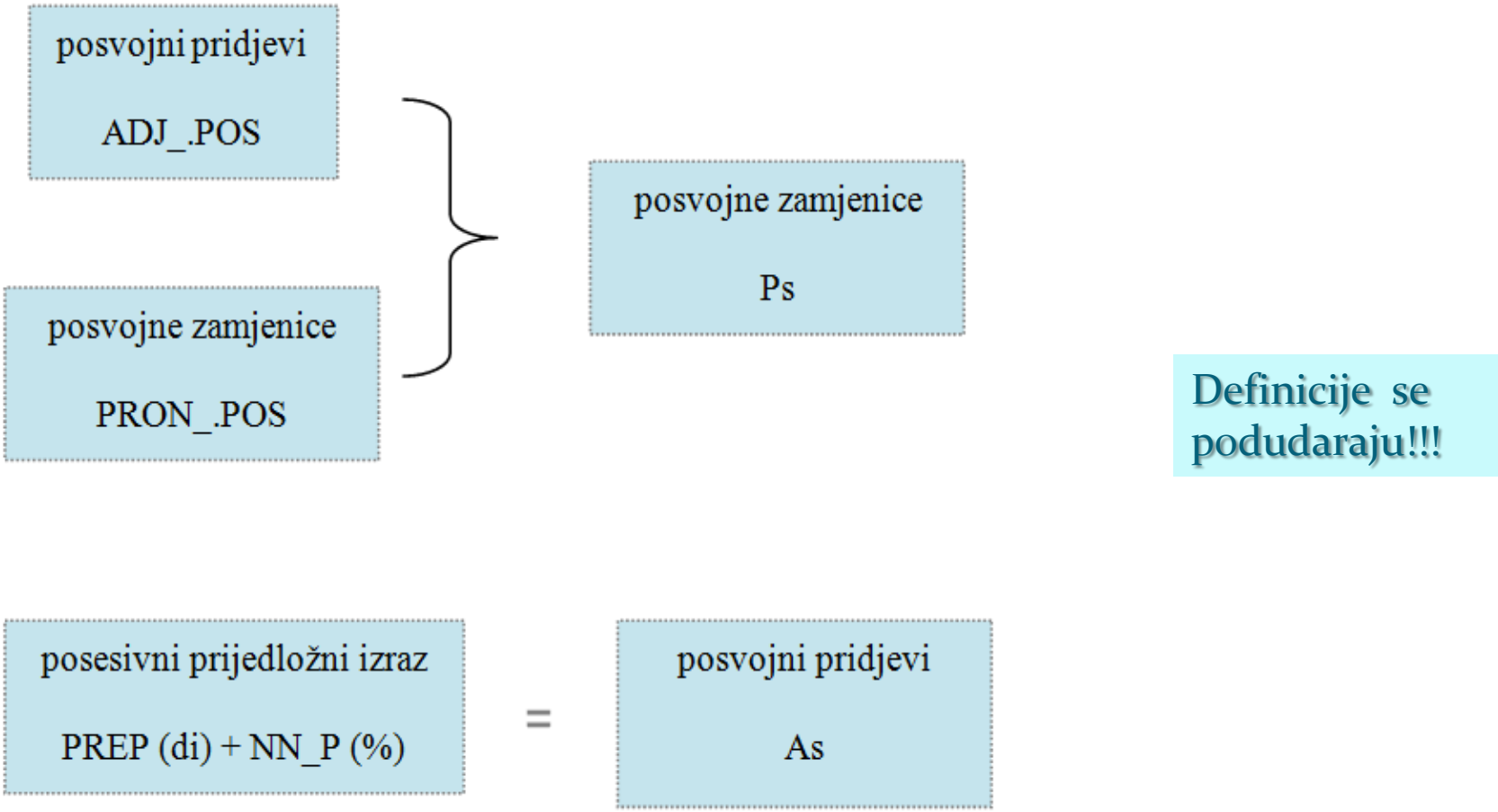
Usklađivanje oznaka III

Razlike u poimanju i postojanju gramatičkih kategorija

itACQUIS			CORIS	HNK	Odabrana oznaka za ovaj rad i objašnjenje	
A			ADJ	A	A	pridjev
As	Ap	An	ADJ	Af	Af	opisni pridjev
			~~ADJ_NUM~~	~~Ao~~	~~Ao~~	~~brojevni pridjev~~[321]
~~APs~~	APp	APn	ADJ_POS	Ps	Psx	posvojna zamjenica/pridjev (it)/povratno-posvojna zamjenica (hr)
			PRON_POS	Px		
P			PRON	P	P	zamjenica
PE			PRON_PER	Pp	Ppx	osobna zamjenica/povratna zamjenica (hr)
PC				Px		
PD			PRON_DIM	Pd	Pd	pokazna zamjenica/pridjev (it)
DD			ADJ_DIM			
PI			PRON_IND	Pi	Piqr	neodređena zamjenica/pridjev (it)/upitna zamjenica/pridjev (it)/odnosna zamjenica/pridjev (it)
DI			ADJ_IND			
PQ			PRON_IES	Pq		
DQ			ADJ_IES			
PR	DR		PRON_REL	Pr		

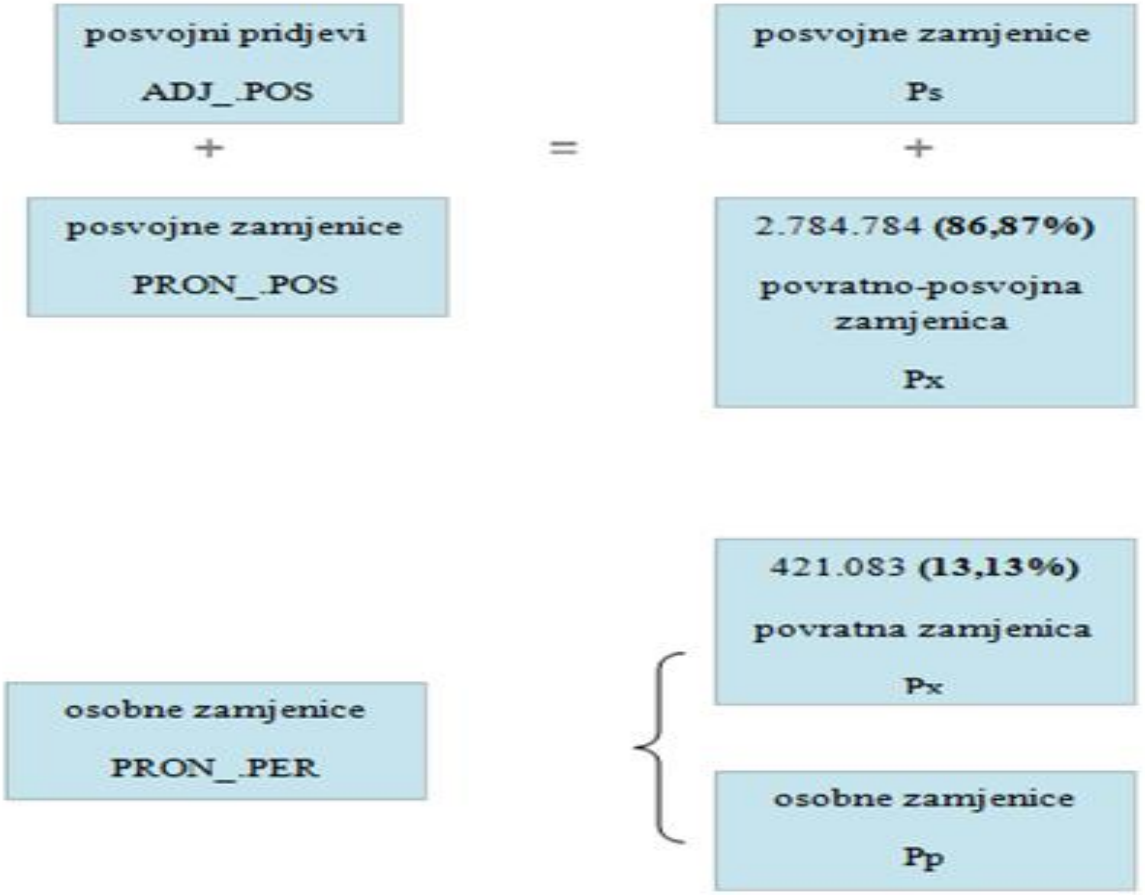
Tablica 3: Prijedlog zajedničkih oznaka za zamjenice i pridjeve

Primjer postupka usklađivanja I



Slika 1. Mogućnosti kontrastivne analize posvojnih zamjenica odnosno pridjeva u postojećim korpusima

Primjer postupka usklađivanja II



1) [msd="Px.*pn.*"]

Pojašnjenje: p→personal n→nominal

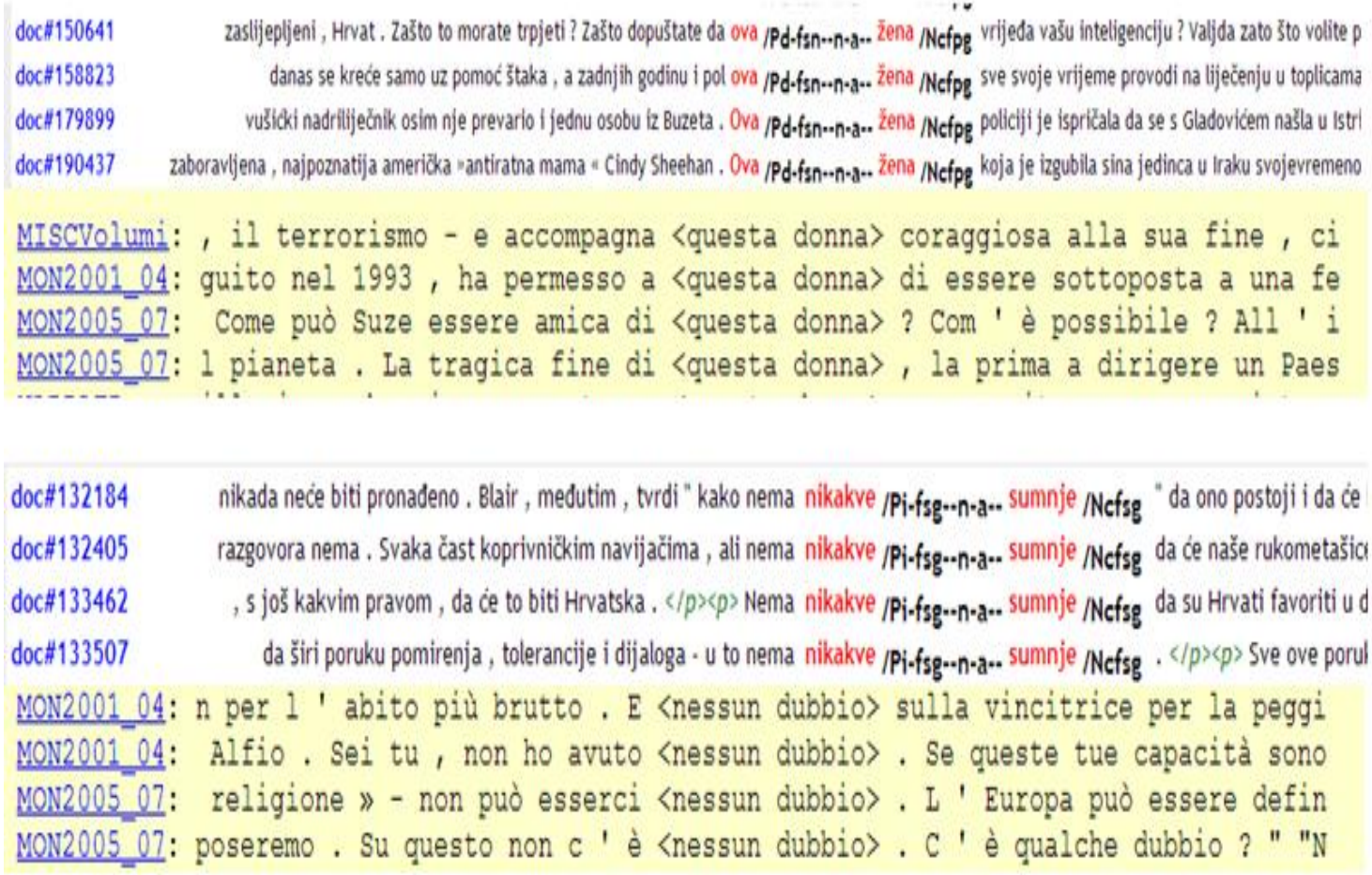
2) [msd="Px.*nsa.*"]

Pojašnjenje: s→possessive a→adjectival

Razlika: tal. povratne zamj. ⟶ lične
Rješenje: odvajanje *se* (Px) od *svoj* (Px)

Slika 2. : Prijedlog usklađivanja oznaka povratne i povratno-posvojne zamjenice

Primjer postupka usklađivanja III



Slika 3 . Primjeri hrvatskih zamjenica odnosno talijanskih pridjeva iz HNK-a v3.0 i CORIS-a

Usklađivanje oznaka IV

Razlike u poimanju i postojanju gramatičkih kategorija !

itACQUIS		CORIS	HNK	Odabrana oznaka za ovaj rad i pojašnjenje	
E		PREP	Sp	Sp	prijedlog
EA		PREP_A			
Cc		CONJ_C	Ccs	Cc	veznik nezavisnosloženih rečenica ili konjunktor
			Ccc		
Cs		CONJ-S	Css	Cs	veznik zavisnosloženih rečenica ili subjunktor
			Csc		
NOs	NOn	ADJ_NUM	Ao	M	brojevi i brojevni pridjevi
			M		
NUM		C_NUM			

Tablica 4: Prijedlog zajedničkih oznaka za prijedloge i brojeve

- različiti status brojeva !!!
- rješenje: odvajanje brojevnih pridj. od ostalih pridj. i njihovo uključivanje pod zajedničku oznaku brojeva

Normalizirani korpus

- **Problem člana**

(Ljubičić 2000, Karlić 2013 i dr. + provjera na usporednome korpusu)

Primjeri neiskazivanja vrijednosti člana u hrvatskome jeziku:

„...u vezi s kvalitativnim Ø značajkama riže...“(jrc31999R0691)
*„...per quanto riguarda **le** caratteristiche qualitative del riso...“(jrc31999R0691)*

„...za potvrđivanje usklađenosti Ø određenog proizvoda ili Ø obitelji proizvoda...“(jrc31999R0691)
*„...se per **un** dato prodotto o **un** gruppo di prodotti determinati...“(jrc31999D0089)*

Primjeri iskazivanja vrijednosti člana u hrvatskome jeziku:

*„...budući da je zato poželjno odrediti **onaj** koncept proizvoda ili obitelji...“(jrc31999D0089)*
*.è opportuno definire **il** concetto di prodotto o di gruppo di prodotti...“(jrc31999D0089)*

*„...«proizvođač» je **svaka** osoba čijom aktivnošću nastaje otpad...“(jrc32006L0012)*
*.«produttore»: **la** persona la cui attività ha prodotto rifiuti...“(jrc32006L0012)*

Normalizirani korpus

- Normalizirana veličina korpusa kao zaključak
  - promatranje distribucije unutar zajedničkih dijelova korpusa
    - samo oznake koje su zajedničke i relevantne za istraživanje

Isključeni:

- ✓ članovi i čestice
- ✓ kratice, pokrate, simboli, interpunkcijski znakovi
- ✓ uzvici

Normalizirani korpus uključuje:

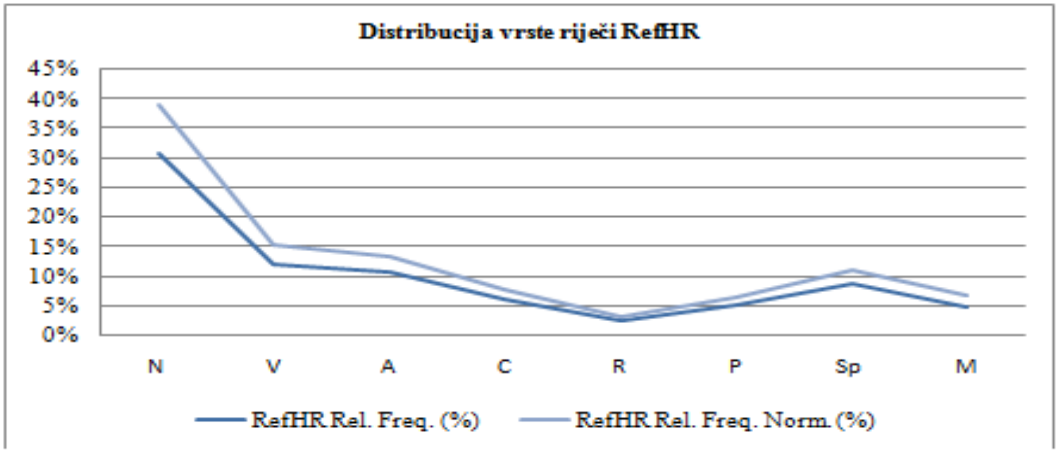
imenice (N), glagole (V), pridjeve (A), veznike (C), priloge (R), zamjenice (P), prijedloge (Sp) i brojeve (M).

Opseg izvornih vs. normaliziranih korpusa

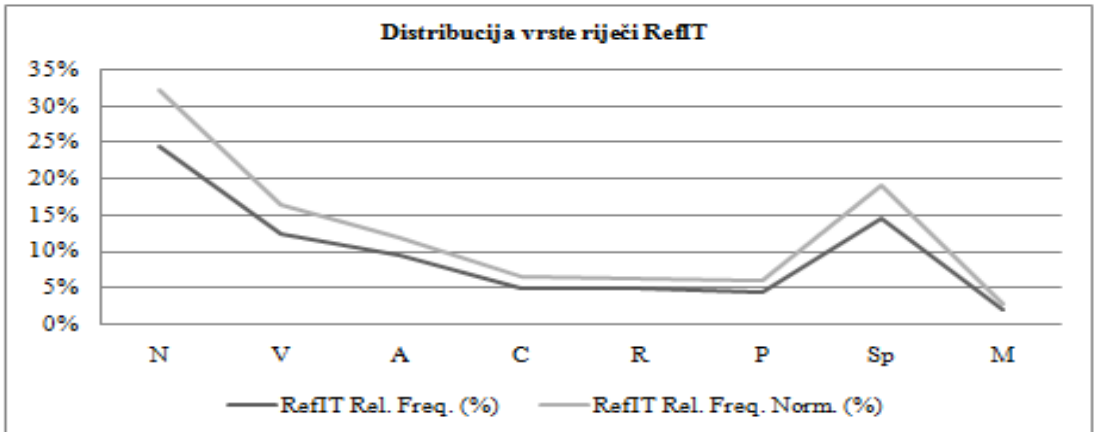
Korpus	Apsolutni broj pojavnica u izvornome korpusu	Apsolutni broj pojavnica u normaliz. korpusu	Omjer broja pojavnica u izvornome i normaliz. korpusu
HNK	216 812 148	177 216 713	0,8173
NN_{1990_2013}	92 363 788	70 852 385	0,7671
hrAcquis	11 209 795	8 784 240	0,7836
CORIS	130 294 347	98 300 670	0,7544
PRGAMM	9 575 784	7 325 921	0,7650
itAcquis	16 955 483	12 464 404	0,7351

Normalizirani vs. nenormalizirani korpusi

Grafikon 1. Usporedba distribucije vrste riječi u RefHR



Grafikon 2. Usporedba distribucije vrste riječi u RefIT



Distribucija vrsta riječi u HNK-u i CORIS-u

Zaključak

- važnost i sustavno planiranje izrade skupa oznaka u skladu sa standardima

  ⮌

  preduvjet usporedivosti među korpusima na unutarjezičnoj i međujezičnoj razini

- hrvatski +, talijanski –

- neusklađenost oznaka, tj. postojanje različitih skupova POS i MSD oznaka za isti jezik
  - onemogućuje unutarjezičnu analizu korpusa i ograničava primjenu

Zaključak

- neizbježno promišljanje i usklađivanje oznaka s obzirom na razlike u poimanju i postojanju gramatičkih kategorija u pojedinim jezicima

- normalizacija korpusa ➡ rješenje za određene analize

  ↪ usporedivost i pouzdanost rezultata

- usporedba i usklađivanje MSD ili POS oznaka - kontrastivna analiza

- budući korpusi ➡ univerzalni skupovi

  ↪ takva usklađivanja neće biti potrebna.