Predstavljanje Balkanološkog instituta SANU na Seminaru za jezičke tehnologije i resurse

Teodora Vuković

19.11.2015.

Tekući projekti i istraživanja

- Jezik, folklor i migracije na Balkanu

TIMSKA TERENSKA ISTRAŽIVANJA

BALKANOLOŠKOG INSTITUTA SANU
(GRAĐA ARHIVIRANA U DABIJU 1999–2015)

Timska terenska istraživanja Balkanološkog instituta SANU (građa arhivirana u DABIju 1999–2015)

Multimedija baza kulturnog nasleđa na Balkanu

- Internet prezentacija građe Balkanološkog instituta
- Moguća pretraživost pomoću interaktivne mape
- Sortirano po temama, koje su vezane za vremenski interval na snimku

Bunjevački rečnik

- Audio rečnik **realne upotrebe jezika**
- Polaziše su **terenski snimci** prikupljeni od strane istraživača sa Balkanološkog instituta tokom 2009. godine
- U sastav rečnika ulaze **lekseme koje se ne koriste u standardnom srpskom jeziku**, ili se koriste ne drugačiji način
- Definicije su preuzete iz **Rečnika bačkih Bunjevaca** Marka Peića i Grga Bačlije
- Obrada podataka pomoću programa za obradu zvuka (SounForge, Audacity)

Bunjevački rečnik

- **disnotor** *im m* 1. svinjokolj, vreme klanja svinja.

Primeri:

1. Svako je za Božić u to vreme zaklo jedno manje svinjče, da Božić opremi. A pravi disnotor, barem naše familije, smo pravili kad je dečiji raspust.

1. Kad uđeš unutra onda čestitaš disnotor, tako se čestita i onda, ovaj, ima, grliš nekog.

Bunjevački korpus

- Model dijalekatskog korpusa bunjevačkog govora
- 50 000 tokena / 45 000 reči
- Normalizacija u 3 nivoa prema standardnom srpskom jeziku
  - Morfofonološke razlike
  - Sintaksičke razlike i prevodive razlike u leksici
  - Neprevodive razlike u leksici
- POS anotacija normalizovanog teksta
- Lematizacija normalizovanog teksta
- Pretraživanje u Corpus Workbenchu pomoću komandne linije

Bunjevački korpus

original	norm1	norm2	norm3
samo	samo	samo	samo
imam	imam	imam	imam
šotošku	šotošku	šotošku	šotošku*
,	,	,	,
pa	pa	pa	pa
kad	kad	kad	kad
ne	ne	ne	ne
sćam	sećam	sećam	sećam
se	se	se	se
ja	ja	ja	ja
na	na	***	***
to	to	toga	toga

ORIGINAL	POS1	LEM1	NORM1	POS2	LEM2	NORM2	NORM3
S	Sp	s	S	Sp	s	S	S
otim	Pd	otaj	tim	Pd	taj	tim	tim
siče	Nc	siča	seče	Vm	seći	seče	seče
snopove	Nc	snopova	snopove	Nc	snop	snopove	snopove
.	#	.	.	#	.	.	.
A	C	a	A	C	a	A	A
sitna	Af	sitan	sitna	Af	sitan	sitna	sitna
pliva	Nc	pliv	pleva	Nc	pleva	pleva	pleva
,	#	,	,	#	,	,	,
prvo	Rg	prvo	prvo	Rg	prvo	prvo	prvo
iđe	Vm	ići	ide	Vm	ići	ide	ide
slama	Vm	slamati	slama	Vm	slamati	slama	slama
iz	Sp	iz	iz	Sp	iz	iz	iz
doba	Nc	doba	doba	Nc	doba	dreša	dreša

Šokački rečnik

- Polazište je rečnik govora Bačkog Brega Marka Ivoševa
- Rečnik je potrebo urediti po leksikografskim standardima
- Konsultovan rečnik
- Postoji mogućnost za dodavanje multimedijalnih sadržaja
- Aplikacija se može primeniti i na druge rečnike

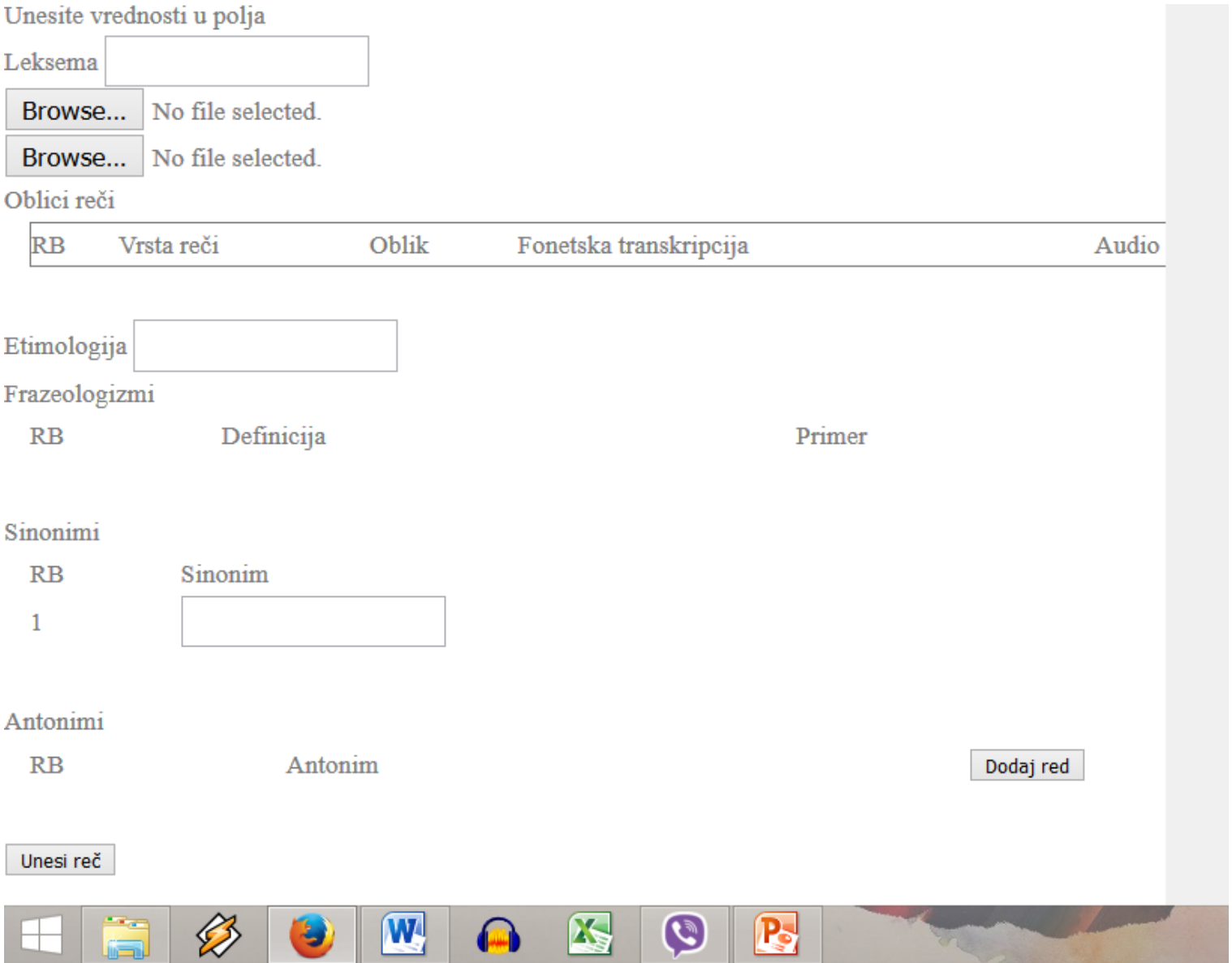
Šokački rečnik

Dodata	Autor	Leksema	Audio i Slika	Etimologija
09/18/2015	svetlana	aljina		

RB	Vrsta reči	Oblik	Fonetska transkripcija	Audio
1	Im ž	aljine		

RB	Definicija	Specijalna leksika	Slika	Audio	Video	Primer
1	ženska jednodelna odeća; haljina				VIDEO	RB / Primeri / Audio
2	odeća uopšte				VIDEO	RB: 1 — Primeri: Jeste l, mamo, sprimili te moje aljine, vrime je da se dižem? — Audio:

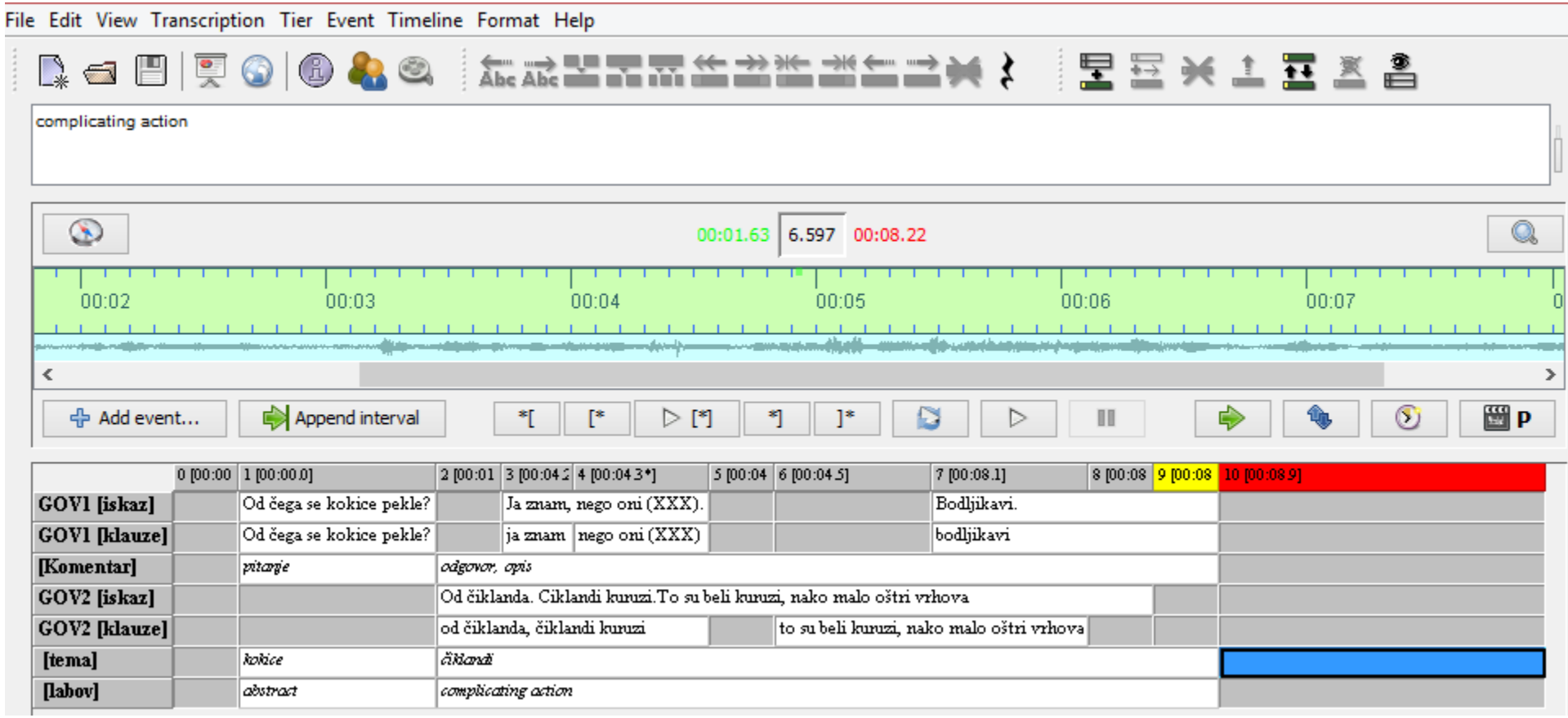
Šokački rečnik



Konstruisanje narativa

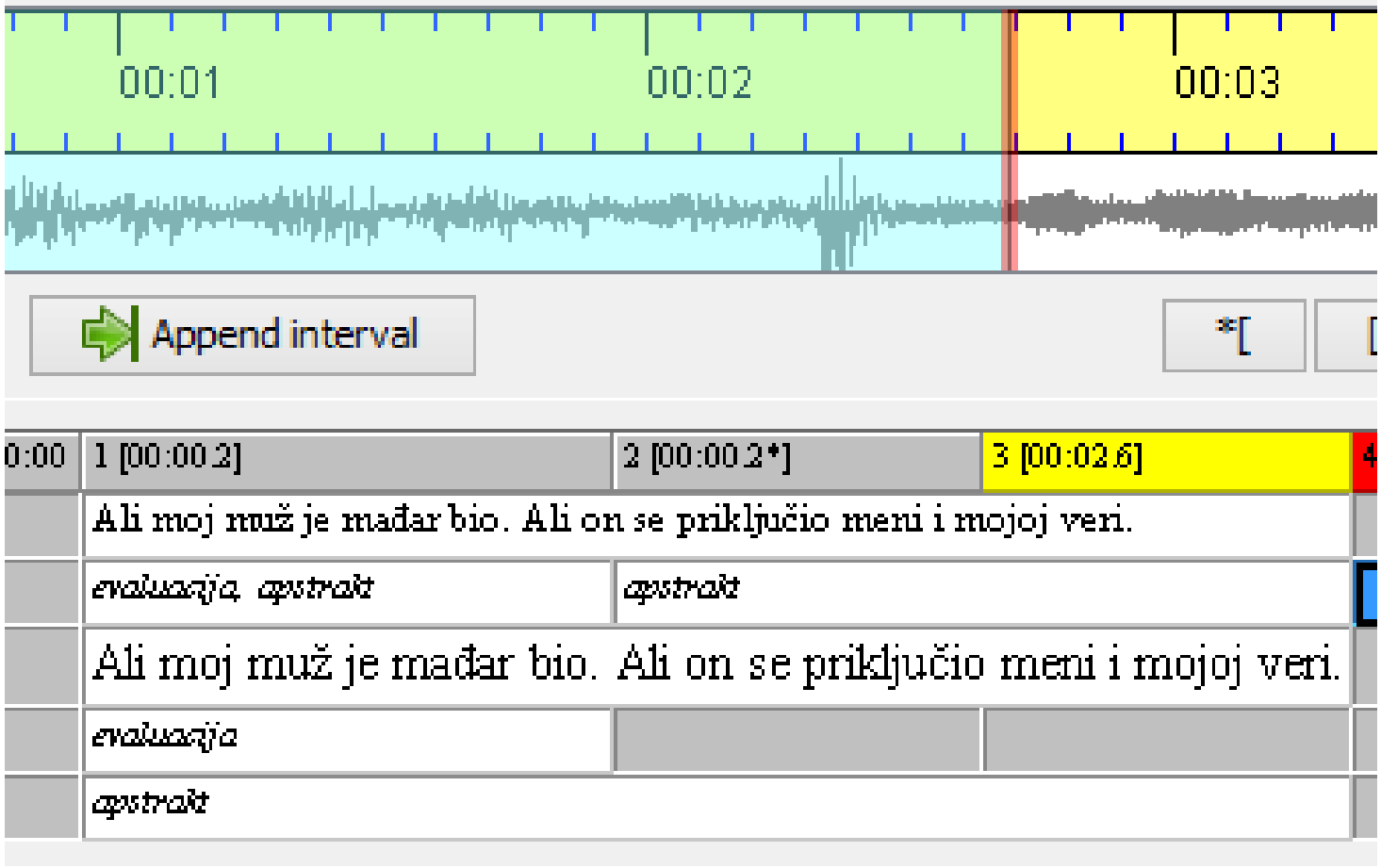
- Projekat u saradnji sa Humbolt univerzitetom u Berlinu
- Upotreba terenskih snimaka iz **srpsko-mađarskih naselja** u okolini Pančeva
- **Transkripcija** u programu **EXMARaLDA**
- Cilj je **korpus narativa** sa anotiranim elementima narativa po značenju i funkciji
- Svrha korpusa **je izučavanje teksta kao niza konstrukcija** (po analogiji sa rečeničnom sintaksom), gde su sastavni elementi konstrukcija, zasebne tematske i funkcionalne celine u okviru teksta

Konsktruisanje narativa



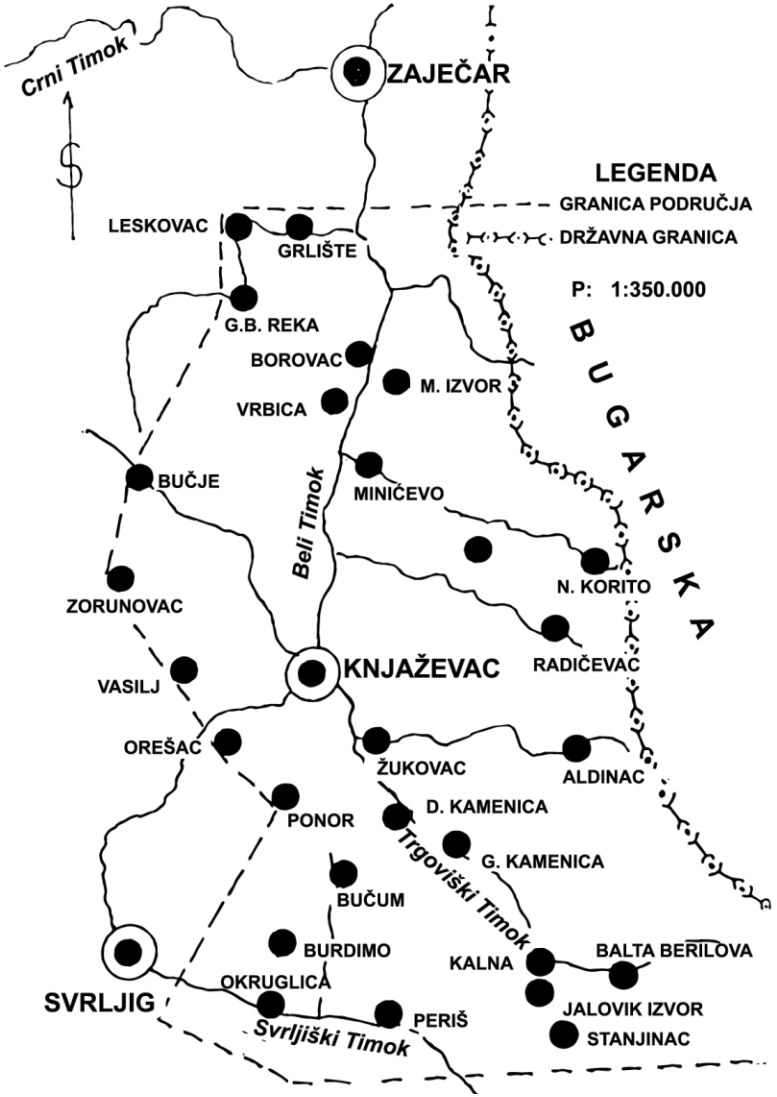
Konstruisanje narativa

нагло умро. Тако да сам ја остала сама. Оставила ме и брат'а (плаче) ... Е тако. Али мој муж је Мад'ар био. Мој је муж Мађар био. Али он се прикључ'ио мени и мојој вери. Тако да он је са



Timočki govori

- Terensko istraživanje u oktobru 2015
- Opštine Knjaževac, Zaječar i Svljig



Timočki govori

- 150+ sati video i audio materijala
- 2000+ strana književnosti na dijalektu
- 100+ sagovornika
- 50+ sela

- Procena veličine – oko 1,5 milion reči

Timočki govori

- **Internet arhiv** sa **interaktivnom mapom**, putem koje se mogu pregledati materijali iz svakog od sela
- Video, audio, fotografije
- Obrada podataka pomoću programa za editovanje audio i video materijala - Audacity, Power Director

Timočki govori

- Transkripcija materijala u programu EXMARaLDA
- Izrada korpusa
- Morfološka i tematska anotacija
- Pretraživost po mestu, polu, obrazovanju, starosti...
- Izrada mape mikrodijalekata
- Kontrastivna analiza jezika u kontaktu

Hvala na pažnji!