

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Digitalizacija jezika i kulture kroz elektronske korpuse: primer timočkih govora

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Mirić, M., Miličević Petrović, M., Ćirković, S. (2021). Digitalizacija jezika i kulture kroz elektronske korpuse: primer timočkih govora. Belgrado : Savez slavističkih društava Srbije - Filološki fakultet [10.18485/mks_dh_skn.2021.1.ch7].

Availability:

This version is available at: <https://hdl.handle.net/11585/886705> since: 2022-06-12

Published:

DOI: http://doi.org/10.18485/mks_dh_skn.2021.1.ch7

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Mirić, M., **M. Miličević Petrović** and S. Ćirković (2021) Digitalizacija jezika i kulture kroz elektronske korpuse: primer timočkih govora. In A. Vraneš (Ed.), *Digitalna humanistika i slovensko kulturno nasleđe*. Beograd: Savez slavističkih društava Srbije i Filološki fakultet. 75-94.

The final published version is available online at:
https://doi.org/10.18485/mks_dh_skn.2021.1.ch7

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

Мирјана Мирић

Балканолошки институт САНУ, Београд

mirjana.miric@bi.sanu.ac.rs

Маја Миличевић Петровић

Филолошки факултет Универзитета у Београду

m.milicevic@fil.bg.ac.rs

Светлана Ћирковић

Балканолошки институт САНУ, Београд

svetlana.cirkovic@bi.sanu.ac.rs

ДИГИТАЛИЗАЦИЈА ЈЕЗИКА И КУЛТУРЕ КРОЗ ЕЛЕКТРОНСКЕ КОРПУСЕ:

ПРИМЕР ТИМОЧКИХ ГОВОРА* **

У раду се представља текући пројекат израде електронског корпуса тимочких говора, као примера дијалекатског корпуса важног за област дигиталне хуманистике. Описују се теренска истраживања у оквиру пројекта *Чувари нематеријалне баштине тимочких говора* (2015–2017) током којих је прикупљена обимна мултимедијална грађа. Приказују се процедуре у изради корпуса: транскрипција и анотација језичке грађе и могућности претраге. Наводе се досадашња истраживања и могуће примене у лингвистичким дисциплинама и областима традицијске културе и фолклорне традиције.

Кључне речи: дигитална хуманистика, електронски корпус, мултимедијална грађа, тимочки говори, дијалектологија, балканска лингвистика, традицијска култура, фолклорна традиција

* Рад представља резултат пројекта *Језик, фолклор и миграције на Балкану* Балканолошког института САНУ (бр. 178010) и *(Dis-)entangling traditions on the Central Balkans: Performance and perception* који се, у оквиру програма ERA.Net RUS Plus, одвија у сарадњи Филолошког факултета Универзитета у Београду, Балканолошког института САНУ у Београду, Семинара за славистику Универзитета у Цириху (Швајцарска) и Института за лингвистичка истраживања Руске академије наука у Санкт Петербургу (Русија). Оба пројекта у Србији финансира Министарство просвете, науке и технолошког развоја Републике Србије.

** Делови рада претходно су изложени на међународним скуповима *XLVIII Международная филологическая научная конференция* (18 – 27. март 2019, Санкт Петербург, Русија) и *Дигитална хуманистика и словенско културно наслеђе* (6 – 7. мај 2019, Београд, Србија).

1. Увод

У раду се описује пројекат израде електронског корпуса тимочких говора, који је у току, и који представља јединствен пример дијалекатског корпуса српског језика чији је циљ дигитализација језички и културолошки релевантне грађе са простора источне Србије. Постојећи корпуси српског језика оријентисани су првенствено на писане текстове из књижевности, новина и уџбеника, писане на стандардном српском језику (в. *Корпус савременог српског језика*, Utvić 2013), или обухватају писану комуникацију која се одвија на интернету (в. *srWaC*, Ljubešić and Klubička 2014; *ReLDI-sr (manually tagged Serbian tweets)*, Miličević i Ljubešić 2016). Корпуси говорног српског језика су изузетно ретки (в. Savić i Polovina 1989 за корпус разговорног српскохрватског језика), а често и посебног карактера, као што је случај са *Српским електронским корпусом раног дечијег говора* (Анђелковић, Шева и Московљевић 2001). На сличну праксу наилазимо и у креирању корпуса других словенских језика, где такође доминирају корпуси стандардних језика (в. преглед у Utvić 2013), премда не изостају ни дијалекатски корпуси (в. нпр. Komrsková et al. 2017 за чешки).

Слаба заступљеност корпуса говорног језика, у које спадају и дијалекатски корпуси, последица је специфичности методологије прикупљања и обраде усмених језичких података. С једне стране, креирање корпуса говорног језика укључује аудио и/или видео-бележење разговора са саговорницима и транскрипцију снимљеног материјала, што прикупљање података чини дуготрајним и организационо и финансијски захтевним. С друге стране, процес формирања електронског корпуса од транскрибованог материјала само се делимично може ослонити на постојеће алате за аутоматску анотацију текста и захтева додатну ручну обраду, будући да су алати углавном креирани за стандардне писане језичке варијете. Имајући у виду наведене специфичности, израда дијалекатских корпуса изузетно је сложена, али уједно и драгоцена.

Захваљујући пракси да се при објављивању дијалектолошких студија дијалекатски текстови углавном наводе у целини, истраживачима српског језика већ је доступна богата дијалекатска грађа. За тимочке говоре у бројним судијама је доступна грађа у штампаном облику (в. Белић 1905; Sobolev 1998; Младеновић 2013, између осталих). Међутим, ако се изузме пилот корпус буњевачког говора (Vuković and Miličević 2017), електронски корпуси дијалеката, који би значајно олакшали приступ и анализу дијалекатске грађе, још увек нису развијени. Циљ овог рада је да прикаже

корак у креирању дијалекатског корпуса тимочких говора и да илуструје могућности примене прикупљеног материјала у областима лингвистике, дијалектологије, балканологије, антрополошке лингвистике и фолклористике.

2. Теренска истраживања

Корпус тимочких говора настаје на основу теренске грађе документоване у оквиру пројекта *Чувари нематеријалне баштине тимочких говора*, који је 2015. и 2016. године финансирало Министарство културе и информисања Републике Србије. Накнадна теренска истраживања су спроведена 2017. године са циљем да се допуни и документује грађа из насеља која у истраживањима претходне две године нису покривена.

2.1. Одабир насеља и саговорника

За потребе пројекта направљена је мрежа пунктова који припадају, у дијалектолошком смислу, тимочким говорима. Тимочки говори припадају тимочко-лужничком дијалекту, као једном од пет дијалеката призренско-тимочке дијалекатске области. Тимочко-лужнички дијалекат обухвата тимочке, белопаланачке, пиротске и лужничке говоре (Собољев 1995: 208; Ивић 2009: 198–199). Као границе тимочких говора за ову прилику узете су оне које помиње Јакша Динић у свом *Тимочком дијалекатском речнику* (2008): на истоку граница тимочких говора је истовремено и граница Србије према Бугарској, на југу развође Тимока и Нишаве и венац Сврљишких планина, на северу Грлишка река. Западна граница иде од Сврљишких планина до изворишног тока Грлишке реке (Динић 2008: IX).¹ У оквирима овако оцртаних граница тимочким говорима припада 96 насеља.² Имајући у виду да је број становника у неким насељима већ у пописној години – 2011 – био мали, документовање теренских разговора је обављано и у другим, већим насељима, са пресељеницима из насеља са данас веома малим бројем становника, или насеља која су данас у потпуности депопулисана.

¹ Границе тимочких говора се помињу и у старијој, једнако релевантној дијалектолошкој литератури (в. Белић 1905; Станојевић 1911).

² Комплетан списак насеља и број становника у њима наведен је у: Ћирковић (2018а: 8–9). Наведена насеља данас припадају трима административним центрима – Књажевцу, Сврљигу и Зајечару.

Од наведених 96 насеља, теренским истраживањима документована је грађа из 92 насеља.³ Теренска документација тимочких говора данас представља мултимедијални корпус грађе који садржи 398 сати аудио-снимака, 192 сата видео-снимака и 4.460 фотографија. У складу са правилима архивирања грађе, корпус грађе је депонован на три места – у Дигиталном архиву Балканолошког института САНУ (Београд), на Филозофском факултету Универзитета у Нишу и у Народној библиотеци „Његош“ у Књажевцу.⁴

Имајући у виду да је приоритет у извођењу пројекта било документовање тимочких говора, у складу са дијалектолошким принципима бирани су саговорници старије животне доби. Насеља која припадају тимочној говорној зони углавном су напуштена или у њима има мало становника, па су управо саговорници најстарије генерације и били доступни истраживачима. Иако би се могло претпоставити да живот у изолованим срединама гарантује очуваност дијалекта или локалног говора, још током истраживања је уочен утицај стандардног језика и губљење неких од проминентних дијалекатских црта, чак и код саговорника за које се претпостављало (узимајући у обзир изолованост и ограничену мобилност саговорника) да су у потпуности дијалектолошки поуздани. Интензивни контакти са говорницима стандардног језика (или варијетета блиских стандарду, које карактеришу одступања од типичних дијалекатских црта), изложеност медијима, само су неки од узрока процеса мењања локалног језичког идиома. Међу саговорницима нашли су се и локални активисти – „чувари традиције“, учитељи, наставници, административни радници, чији су наративи драгоцени за истраживања усмене историје, фолклорне традиције, етнографије, али дијалектолошки нису поуздани. Ипак, прикупљена усмена грађа од саговорника оваквог типа може бити драгоцен за социolingвистичка истраживања, и то она која прате скалу трансформација дијалекатских црта. Накнадно су, током 2018. године, обављена истраживања са млађим саговорницима средњошколског узраста, како би се обезбедио минималан узорак говора млађе генерације.

2.2. Примењена методологија истраживања

³ Организациону и логистичку подршку истраживачком тиму су обезбедили сарадници Народне библиотеке „Његош“ у Књажевцу.

⁴ Интерактивна мапа истражених насеља доступна је на сајту <http://balksrv2012.sanu.ac.rs/webdict/timok/index> (приступљено 20.5.2019), а на Јутјуб каналу „Теренска истраживања“ (<https://www.youtube.com/channel/UC4EpCSAnEb2RIsIRY7pfNdQ>, приступ 8. јуна 2019) налази се велики број монтираних теренских аудио и видео-записа.

У документовању тимочких говора нису коришћени лексички, термилошки, морфосинтаксички упитници, већ је у истраживањима примењена методологија отвореног интервјуа, данас често коришћена у антрополошко-лингвистичким истраживањима.⁵ Имајући у виду да је циљ истраживања осим лингвистичког документовања био и документовање традицијске културе и фолклорне традиције, у разговорима је коришћен оквирни упитник који је покрио екстензиван списак тема за разговор. Циљ овако конципиране методе истраживања није био да се у једном разговору покрију све оквирним упитником наведене (или предвиђене) теме,⁶ већ да се сам разговор прилагоди саговорнику, те да се елицитира дужи наратив са што мање потпитања истраживача. Истраживачки тим нису у целини чинили изворни говорници локалних тимочких говора (неки од истраживача нису носиоци ниједног варијетета који припада призренско-тимочкој дијалекатској области) и примарни циљ овако конципираних истраживања био је да се избегне утицај истраживачевог идиома на саговорников идиом, или нивелација говора саговорника према говору истраживача.⁷

У теренским истраживањима примењивана је до сада најчешће практикована метода аудио-бележења разговора, али је у овом пројекту примењено и видео-бележење. И за аудио и за видео-бележење од саговорника је тражена сагласност, која је забележена на аудио-снимку. Кад год је за то било могућности, разговори су вођени у затвореним просторијама, како би се у највећој могућој мери смањила бука, како би сам саговорник био што чујнији. Видео-документовање разговора у овом пројекту је примењивано ради потенцијалних истраживања невербалне комуникације, као и визуализације контекста у којем је вођен теренски разговор.

Мултимедијални теренски корпус грађе, који се састоји од аудио и видео-записа, фотографија, садржаја који покрива различите теме како из традицијске културе, усмене историје, (ауто)биографских прича, испричаних на локалним идиомима, или варијетету блиском стандардном језику, даје простор за дисциплинарно различита истраживања и научне анализе, као и за интердисциплинарна и мултидисциплинарна истраживања.

⁵ О развоју истраживачке методе у антрополошко-лингвистичким истраживањима у Србији видети више у: Сикимић (2012).

⁶ О оквирном упитнику и темама разговора видети више у: Ћирковић 2018а.

⁷ О односу истраживача и саговорника као комуникацијског пара видети више у: Norrick (2003); Сикимић (2004); Ћирковић (2015).

3. Ка електронском корпусу: транскрипција грађе, анотација, претрага⁸

Како би се богата грађа прикупљена током теренских истраживања учинила доступном већем броју истраживача и како би се олакшало њено претраживање, у оквиру међународног пројекта *(Dis-)entangling traditions on the Central Balkans: Performance and perception* приступљено је изради електронског корпуса тимочких говора. Основу за израду корпуса представљају аудио и видео-записи усмене продукције носилаца тимочких говора.

3.1. Избор и транскрипција прикупљене грађе

У уводном делу рада већ је истакнуто да израда корпуса говорног језика представља врло сложен задатак који чини низ међусобно повезаних корака. У случају корпуса тимочких говора, прво је било потребно одабрати снимке који су дијалекатски најрелевантнији.⁹ У ту сврху узето је у обзир присуство неколико кључних одлика тимочких говора, при чему за избор снимка није било неопходно да све буду забележене. На нивоу изговора, тражени су уједначен експираторни акценат, присуство вокалног *л* (*влна*) и полугласа *ь* (*сьг „сад”*), као и одсуство гласа *х* (*леб*). У домену морфологије, у центру пажње су били падежни систем заснован на номинативу и општем падежу (*у школу, из школу*), постпозитивни заменички датив у функцији присвојног придева (*дете ми*), компаратив изведен помоћу префикса *по-* (*високо>повише*) и одсуство инфинитива глагола. Паралелно са дијалекатским одликама, тежило се и постизању што уједначеније територијалне распрострањености. Услед недостатка релевантних говорника, није било могуће укључити у корпус продукцију из свих, већ из 63 од укупно 92 места (в. Vuković submitted). Као циљна величина корпуса одређено је 500.000 речи.

Наредни корак чинило је пребацивање снимака у текстуални формат. Транскрипција говорног језика позната је у корпусној лингвистици као веома тежак задатак чак и када је у питању стандардни језик, а мање конвенционални корпуси попут дијалекатских намећу додатне изазове везане за разумевање изговореног и начин бележења нестандартних одлика, али и за различите потребе истраживача (в. радове у

⁸ Опис процедуре израде корпуса делимично је заснован на Вуковић и Самарџић (2018) и Vuković (submitted).

⁹ Као што је напоменуто у претходном одељку, у теренским разговорима забележен је и сразмерно висок удео продукције на стандардном језику или варијетету блиском стандарду.

Beal, Corrigan and Moisl 2007). За транскрипцију снимака тимочких говора ангажована је група сарадника чији рад је био заснован на детаљним смерницама (Vuković 2016). Један од главних принципа у смерницама јесте примена полуортографске транскрипције. У полуортографској транскрипцији тежи се што вернијем бележењу изговореног, али уз употребу стандардног писма, у овом случају латиничног. Посебним симболима су у корпусу бележени гласови којих нема у стандардном српском језику, попут полугласа или палатализованог *к* (за детаље в. Vuković submitted), а додато је и неколико посебних ознака везаних за особине говорног језика (пре свега „средња тачка” · за означавање паузе и # за нејасне делове у снимцима). Међу осталим критеријумима значајни су: записивање искључиво малим словима, уз изузетак употребе великих слова за означавање наглашених слогова (нпр. *živIm, ovOj*), бележење дугих вокала понављањем слова (*Aaako*), означавање недовршене речи косом цртом (нпр. *kupU/<kupUva*), као и означавање преклапања говорника средњим заградама. Завршене транскрипте прегледају и по потреби исправљају чланови истраживачког тима.

Транскрипција је спроведена у програму *Partitur-Editor* из софтверског пакета *EXMARaLDA* (Schmidt 2009).¹⁰ Овај програм омогућује одвајање продукције различитих говорника (уколико их у снимку учествује више) у засебне нивое у транскрипту, као и сегментацију унутар појединачних нивоа. Сегментација говорног језика увек представља велик изазов. Смернице за корпус тимочких говора упућују на интуитивно издвајање краћих интонацијско-смисаоних сегмената, без обзира на то да ли чине синтаксичку целину или не. Трајање већине издвојених сегмената креће се између једне и четири секунде.

Као крајњи производ транскрипције, добијају се документи у формату *xml*. Овај формат транскрипта уз основни текст чува информације о говорницима, подели на исказе и времену када су се различити делови разговора одиграли (што ће омогућити да у корпусу текст буде поравнат са снимком).

Паралелно са транскрипцијом врши се и унос метаподатака, који обухватају податаке о говорницима (њиховој животној доби, полу, години рођења, месту становања, месту порекла, нивоу образовања и занимању)¹¹ и о снимцима (локацији на којој су снимљени, географским координатама те локације, регији, месту снимања

¹⁰ <https://exmaralda.org/en/> (приступ 7. јуна 2019)

¹¹ Треба напоменути да подаци о животној доби, години рођења и нивоу образовања не постоје за сваког саговорника, будући да су неки истраживачи у интервјуима примењивали антрополошку методу у којој се од саговорника не тражи да наведу личне податке попут имена и годишта.

разговора (кућа, двориште и сл.), темама о којима се говори). Унети метаподаци ће будућим корисницима омогућити креирање поткорпуса и претраживање, на пример, само транскрипата из одређених места, или само оних у којима се говори о одређеној теми.

3.2. Лингвистичка анотација транскрипата

Како би се омогућили напредни видови претраге, у транскрибован текст уносе се додатне лингвистичке информације, односно врши се аотирање текстова. Унос додатних информација обухвата три нивоа: морфосинтаксичко означавање, лематизацију и нормализацију. Морфосинтаксичка анотација подразумева додељивање свакој речи ознаке врсте речи и других релевантних граматичких категорија (броја, падежа, времена, и сл.). Лематизација представља прикључивање свакој речи њеног основног облика, који одговара речничкој одредници (на пример, номинативу јединине за именице). Најзад, нормализацију чини придруживање одговарајућих стандардних облика (на пример, *kipil* > *kipio*), при чему облици који већ одговарају стандарду на овом нивоу анотације остају непромењени.

Морфосинтаксичка анотација и лематизација данас су присутне у већини корпуса, захваљујући доступности алата за њихово аутоматско спровођење. Међутим, у случају дијалеката непоклапања са стандардним језиком спречавају искључиву примену постојећих (стандардних) алата. Стога је за корпус тимочких говора одабран приступ који чине следеће три компоненте: 1) означавање мањег узорка транскрипата алатима за стандардни српски језик, 2) ручна провера и исправљање добијених ознака, 3) тренирање (на исправљеним ознакама), помоћу метода машинског учења, посебних аутоматских алата за тимочке говоре.

Комбиновано аутоматско и ручно означавање спроведено је на узорку од приближно 20.000 речи користећи алате које су развили Ljubešić et al. (2016).¹² Ови алати користе морфосинтаксичке ознаке дефинисане у оквиру иницијативе MULTEXT-East (верзија 5).¹³ Свака ознака састоји се из низа алфанумеричких карактера који представљају вредности релевантних категорија, које се кодирају на унапред одређеним позицијама. На пример, *Pp3fsd* означава личну (p) заменицу (P) 3. лица (3),

¹² <https://github.com/clarinsi/reldi-tagger> (приступ 7. јуна 2019)

¹³ Преглед ознака за различите језике може се видети путем странице <http://nl.ijs.si/ME/V5/msd/html/> (приступ 7. јуна 2019).

женског рода (*f*) једине (*s*), у дативу (*d*). Међутим, будући да се у тимочким говорима користи постпозитивни члан, који у стандардном српском језику није присутан, током фазе ручне провере било је потребно увести додатне ознаке, *v*, *t* и *n*, које кодирају различите деиктичке облике члана, паралелне облицима стандардних показних заменица (детаљније у Вуковић и Самарџић 2018).

Код лематизације је у сврху упоредивости са другим корпусима, али и у сврху олакшавања претраге, одлучено да се за речи које имају еквивалент у стандардном српском језику користи стандардни основни облик (на пример, *хлеб*, уместо дијалекатског *леб*), док се изворни облик задржава само за дијалекатски специфичне речи попут релативне заменице *кво*. Посебно осетљиво у овом контексту било је питање инфинитива, који се у тимочким говорима не користи. Будући да се у транскриптима јављају и елементи стандардног језика, решено је да се свим глаголима као основни облик ипак додели облик инфинитива.

По завршетку провере и исправљања морфосинтаксичких ознака и лема у узорку, нагласак ће бити стављен на процес тренирања наменских алата за морфосинтаксичку анотацију тимочких говора. Када буде постигнут задовољавајући степен тачности ових алата, они ће бити примењени на преостале транскрипте.

Упоредо са радом на морфосинтаксичкој анотацији, започета је и фаза ручне нормализације одвојеног узорка од 20.000 речи, у који су укључени мањи узорци из већег броја места, како би се обухватила локална варијација (детаљније у Vuković submitted). Нормализација је подељена у два одвојена нивоа, морфо(фоно)лошки и синтаксички.¹⁴ У оквиру првог нивоа, интервенисано је само уколико употребљени облик не одговара додељеној леми, односно не постоји у стандардном српском језику (на пример, *kupil* > *kupio*), док је за други ниво било значајно да ли је употребљени облик граматичан (гледано из перспективе стандардног језика) у датом контексту (*[iz] školu* > *[iz] škole*). Слично процедури за морфосинтаксичку анотацију, крајњи циљ је развој алата за аутоматску нормализацију, на основу ручно нормализованог узорка.

Као резиме лингвистичких информација које се у корпусу придружују изворном тексту, у табели 1 илуструјемо пуну анотацију једног сегмента, у вертикалном формату.

¹⁴ Лексичка нормализација се не врши.

Изворни облик	Лема	Морфосинтаксичка ознака	Морфо(фоно)лошка нормализација	Синтаксичка нормализација
u	u	Sa	u	u
bUgarsku	bugarska	Npfsan	bugarsku	bugarskoj
bIl	biti	Vap-sm	bio	bio
vAšar	vašar	Ncmsnn	vašar	vašar
takAj	takav	Pd-msn	takav	takav

Табела 1. Пример аотираног сегмента.

3.3. Могућности употребе

Након развоја потребних алата и након завршетка аотације свих транскрипата, корпус тимочких говора биће припремљен за укључивање у платформу *NoSketch Engine* (Rychlý 2007),¹⁵ у оквиру репозиторијума словеначког огранка европске инфраструктуре CLARIN.¹⁶ Ова платформа пружа бројне могућности за сложеније претраге, захваљујући опцијама за креирање поткорпуса заснованих на унетим метаподацима, као и имплементацији регуларних израза и језика за претрагу корпуса (*Corpus Query Language - CQL*; в. Jakubíček et al. 2010). На пример, биће могуће једним упитом добити дијалекатске и стандардне облике у продукцији носилаца тимочких говора (тако би резултати за *bio* садржали и облик *bio* и облик *bil*), што може значајно олакшати утврђивање степена присуства дијалекатских облика.

Из корпуса ће додатно моћи да се добију подаци о учесталости појављивања тражених речи, вишечланих израза, морфолошких облика, па и појединих синтаксичких конструкција, што ће олакшати квантитативне анализе тимочких говора. Најзад, електронски корпус тимочких говора ће својим метаподацима и могућностима претраге допринети истраживању културе кроз приступе који су инспирисани рачунарским методама анализе сентимента (на пример, у сврху испитивања ставова тимочких говорника према говорницима сродних дијалеката у Бугарској) и дистрибуционе семантике (на пример, у сврху издвајања кључних појмова који се везују за одређене теме).

4. Лингвистичка истраживања на корпусу

¹⁵ <https://nlp.fi.muni.cz/trac/noske> (приступ 7. јуна 2019)

¹⁶ <http://www.clarin.si> (приступ 7. јуна 2019)

Будући да је израда електронског корпуса тимочких говора у току, досадашња лингвистичка истраживања рађена су или на пробним електронским верзијама корпуса (в. нпр. Вуковић и Самарџић 2018; Vuković submitted) или су аутори студија сами транскрибовали грађу и ексцерпирали релевантне примере (в. нпр. Мирић 2017; Ристић 2018; Сикимић 2018; Ћирковић 2018б). Истраживања обухватају широк спектар лингвистички релевантних тема и крећу се, углавном, у оквирима дијалектологије и балканске лингвистике, а све је присутнија примена квантитативних метода у анализи грађе.

На лингвистичком нивоу истражују се идиолекти саговорника који су изворни носиоци тимочких говора (в. Ђорђевић Пејовић и Ристић 2017 за узорак говора из насеља Локва; Ристић 2018 за Ново Корито; Сикимић 2018 за Заграђе). Аутори исцрпно анализирају одлике говора по језичким нивоима, фокусирајући се на фонетско-фонолошки, морфолошки и лексички ниво. Поменуте студије указују на присуство појединих дијалекатских црта које се сматрају дистинктивним обележјима тимочких говора, мада посебно наглашавају недоследност у њиховој употреби, сматрајући је „прилагођавањима идиолекта према очекиваном 'стандардном' говору” (Сикимић 2018: 149). На обимном електронском корпусу биће могуће брже, аутоматском или полуаутоматском претрагом, анализирати степен присуства дијалекатских црта код појединачних говорника и у дијалекту као целини, како би се сагледало у којој мери и под којим условима се дијалекатске црте чувају, а када и зашто долази до прилагођавања стандардном језичком варијетету.

Осим поменутог типа студија, објављено је или је у току неколико истраживања која на већем узорку саговорника квантитативно анализирају специфичне дијалекатске феномене, махом у домену морфологије и синтаксе. Један део радова посвећен је анализи постпозитивног члана, из перспективе његове просторне дистрибуције (Вуковић и Самарџић 2018). У радовима у области морфосинтаксе указује се на доследно одсуство инфинитива, граматикализацију маркера будућег времена *че/ће* (Мирић 2017), као и на факултативност у употреби субјунктивног везника *да* у конструкцији футура првог (Мирић 2018). Посебна пажња посвећена је истраживањима дативске енклитике *си*, која се осим из перспективе утицаја просторних и друштвених фактора на њену дистрибуцију (Makarova and Vuković, in preparation), анализира и из синтаксичке и семантичке перспективе (Ćirković 2019).

У поменутих радовима језичке црте у тимочким говорима посматрају се најчешће у контексту одлика Балканског језичког савеза. Наиме, говори признанско-тимочке дијалекатске области сматрају се „високобалканизованим српским народним говорима” (Милорадовић 2015: 269) и припадају периферним говорима Балканског језичког савеза. Електронски корпус ће омогућити да се испитају и други балканизми, као што су удвајање објекта, аналитичка деклинација, и други, чија је употреба забележена приликом теренских истраживања.

Електронски корпус тимочких говора биће могуће употребити за језичко профилирање носиоца дијалекта. Наиме, израчунавањем фреквенце присуства проминентних дијалекатских црта, може се утврдити у ком степену је саговорник дијалекатски релевантан (в. више у Конџ, Макарова и Соболев 2019). С друге стране, електронски корпус је значајан за испитивање индивидуалних разлика међу саговорницима, будући да садржи узорке говора више десетина саговорника. Досадашња истраживања указују на висок степен варијације међу саговорницима, која је узрокована различитим друштвеним и ареалним факторима.

Када је реч о ареалним факторима, у електронском корпусу сваком транскрипту придружују се метаподаци о географским координатама, који омогућавају истраживања у домену дијалектометрије и ареалне лингвистике. Наиме, ареални фактори утичу на језичку/дијалекатску варијацију: говор становника оних насеља која су због рељефа терена изолованија, или удаљенија од административних центара, отпорнији је на утицаје других језичких варијета попут стандардног језика, те склонији очувању типичних дијалекатских карактеристика (в. радове у De Busser and LaPolla 2015). Прва истраживања у овом домену на грађи тимочких говора показују значајан утицај надморске висине на очување постпозитивног члана (Вуковић и Самарџић 2018), као и значајан ефекат географске ширине, ареала и насеља на разлику између тимочких и лужничких говора у изостављању субјунктивног везника *да* у конструкцији футура првог (Милић 2019).

У погледу друштвених фактора, електронски корпус отвара могућност и за истраживања утицаја демографских фактора на језичку/дијалекатску варијацију, будући да садржи метаподатке о полу саговорника и животној доби. Ранија истраживања су показала да је говор старијих људи конзервативнији (в. радове у Chambers, Trudgill and Schilling-Estes 2003), што сугерише да се у говору старијих саговорника чувају типичне дијалекатске црте, док млађе генерације употребљавају варијетет близак стандардном језику. С обзиром на то да корпус садржи и узорке

говора млађе генерације саговорника средњошколског узраста, ове наводе могуће је и емпиријски проверити.

5. Истраживања традицијске културе и фолклорне традиције тимочких говора: реализована и потенцијална истраживања

Осим за лингвистичка истраживања, грађа прикупљена у наведеним теренским истраживањима до сада је коришћена, а недвосмислено је погодна и за потребе истраживања из области фолклористике, антрополошке лингвистике, етнолингвистике, социоллингвистике.

Фолклористичке студије објављене на основу документоване тимочке теренске грађе указују на елементе усменог стваралаштва који нестају, али и на оне који опстају и трансформишу се, сагледавају се контекстуални оквири у којима се различити фолклорни жанрови реализују, као и сижејне и жанровске карактеристике усмених записа, теренски записи се пореде да постојећом фолклорном и етнографском грађом и садржајима са различитих интернет портала, и објављује се разноврстан репертоар различитих фолклорних жанрова (Ђорђевић Белић 2018а; 2018б; Ђорђевић Пејовић и Ристић 2017; Ђорђевић Пејовић 2018; Вујновић 2018; Гудураш 2018; Станковић 2018; Поповић Николић 2018). До сада објављене студије указују и на могућности антрополошко-лингвистичког читања теренских наратива, илуструјући традицијску културу и њене трансформације, утицај социо-економских и еколошких фактора на свакодневни живот становника руралних заједница (Сикимић 2017; Сикимић 2018; Ћирковић 2019; Ћирковић, in press). Такође, досадашња истраживања теренски документоване грађе у тимочким насељима су се кретала у правцу истраживања невербалне комуникације, указујући на повезаност гестова и невербалних елемената исказа, као и на олакшано разумевање наратива уз визуелну илустрацију појединих његових сегмената (Ћирковић 2017; 2018б).

Најзад, у креирању корпуса предвиђена је анотација према „макротемама”, односно темама које су наведене у оквирном теренском упитнику, а тичу се традицијске културе (на пример, свадбени ритуали, ритуали везани за рођење детета, посмртни ритуали, сегменти календарског циклуса), а изоставља се анотација мањих сегмената одређених ритуала и њихових елемената. Теренски разговори имају веома сложену структуру, у једном разговору се преплићу тематски различити наративи, па се могу очекивати различити проблеми приликом тематске анотације корпуса. Ипак,

тематска анотација корпуса ће знатно олакшати претрагу корпуса потребну истраживачима који се баве традицијском културом и фолклорном традицијом.

5. Закључна разматрања

У раду је описана грађа прикупљена током низа теренских истраживања усмерених на тимочке говоре и културу, при чему је посебна пажња посвећена текућој изради електронског корпуса заснованог на тој грађи. Дат је преглед већ спроведених лингвистичких, фолклористичких и сродних истраживања, а указано је и на могућности које ће постати доступне након завршетка и објављивања корпуса.

Међутим, електронски корпус тимочких говора неће представљати само драгоцену помоћ у истраживањима, већ и средство за чување нематеријалне, језичке и културне баштине кроз њихову дигитализацију. Подаци добијени у теренским истраживањима сведоче о смањењу броја говорника тимочких говора, с једне стране узрокованог све мањим бројем становника у селима, а с друге стране све приметнијим прилагођавањима говорника варијетету блиском стандардном језику. Депопулација села, миграције становништа из села у град, утицај савремене културе и медија рефлектују се и на традицијску културу, чији се елементи под утицајем наведених околности трансформишу и губе. Самим тим, овај дијалекатски корпус посебно је значајан за дигиталну хуманистику. Наиме, електронски корпус тимочких говора не само да представља дигитализовану верзију теренски прикупљеног дијалекатског материјала, већ је истовремено и дигитална збирка традицијске културе – оних ритуала који се и даље практикују и оних који се данас могу реконструисати само на основу сећања. Чини се да су управо корпуси попут корпуса тимочких говора кандидат за улогу споне између корпусне лингвистике и дигиталне хуманистике, будући да су више усмерени на димензију истраживања традицијске културе од других врста корпуса.

Повезаност корпуса и дигиталне хуманистике потребно је овде посебно нагласити, будући да корпусна лингвистика – упркос врло наглашеној дигиталној компоненти – не налази увек место у оквирима дигиталне хуманистике. Jensen (2014) у свом прегледу односа између ове две дисциплине истиче чињеницу да је корпусна лингвистика дигитално оријентисана од тренутка свога настанка (у савременом смислу термина, који обухвата израду и анализу електронских корпуса) и да би већ самим тим морала више бити узимана у обзир у дигиталној хуманистици. Разлике које уочава

између ових области тичу се степена структурирања података (при чему корпусна лингвистика типично барата структуриранијим подацима) и метода анализе (које су у корпусној лингвистици највећим делом квантитативне, а у дигиталној хуманистици често и квалитативне), али слажемо се са овим аутором у тврдњи да корпусна лингвистика и дигитална хуманистика ипак имају много тога заједничког, почевши од тежње да објасне аспекте људског искуства и људског понашања уз помоћ дигиталних технологија.

Извори и литература

- Анђелковић, Даринка, Нада Шева, и Јасмина Московљевић. *Српски електронски корпус раног дечијег говора*. Доступно на: <https://chilides.talkbank.org/access/Slavic/Serbian/SCECL.html> Веб. 01. 06. 2019.
- Белић, Александар. „Дијалекти источне и јужне Србије“. *Српски дијалектолошки зборник I* (1905). Штампано.
- Вујновић, Татјана. „’Ако их не дираш неће ниједна да те удави’. Теренска истраживања Стогазовца“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 31–42. Штампано.
- Вуковић, Теодора, и Тања Самарцић. „Просторна расподела фреквенције постпозитивног члана у тимочком говору“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 181–199. Штампано.
- Гудураш, Јелена. „’Нови пирати иду са апарати’. Разговор о скривеном благу на Старој планини“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 43–55. Штампано.
- Динић, Јакша. *Тимочки дијалекатски речник*. Београд: Институт за српски језик САНУ, 2008. Штампано.
- Ђорђевић Белић, Смиљана. „Здравица: (традиционални?) жанр и савремени контексти“. *Књижевна историја* 164 (2018а): 71–106. Штампано.

- Ђорђевић Белић, Смиљана. „Локална култура и стваралаштво Милоша Петровића из Васиља“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018б. 65–98. Штампано.
- Ђорђевић Пејовић, Сузана. „Тој ти сад причам, а ти да знаеш, па некад да причаш на твоју децу.“ Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 57–64. Штампано.
- Ђорђевић Пејовић, Сузана, и Бојан Ристић. „Водени бик у изворима из Локве: Фолклористички и дијалектолошки осврт“. Ђорђина Трубарац Матић (ур.). *Водени бикови и водене краве у усменим традицијама света, Фолклористика 2/2* (2017): 47–61. Штампано.
- Ивић, Павле. *Српски дијалекти и њихова класификација*. Слободан Реметић (прир.). Нови Сад, Сремски Карловци: Издавачка књижарница Зорана Стојановића, 2009. Штампано.
- Конџер, Дария В., Анастасия Л. Макарова, и Андрей Н. Соболев. „Статистический метод языкового профилирования носителя диалекта (на материале восточносербского идиома села Берчиновац)“. *Вестник Томского государственного университета. Филология* 58 (2019): 17–33. Штампано.
- Милорадовић, Софија. „Српски периферни говори – међујезички утицаји и балканистички процеси“. *Исходишта I* (2015): 267–279. Штампано.
- Мирјић, Мирјана. „Степен граматикализације футура првог у тимочким говорима“. *Зборник Матице српске за филологију и лингвистику LX/1* (2017): 133–164. Штампано.
- Мирјић, Мирјана. „Употреба/изостављање субјунктивног маркера да у конструкцији футура првог у тимочким говорима“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 201–218. Штампано.
- Младеновић, Радивоје. *Говор јужнокосовског села Гатње*, Београд: Институт за српски језик САНУ, 2013. Штампано.
- Поповић Николић, Данијела. „’Сваке приче се не причају код свакога’. Избор из усмених прозних облика Дебелице“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац:

- Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 119–125. Штампано.
- Ристић, Бојан. „Ново Корито: Село на граници“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 127–144. Штампано.
- Сикимић, Биљана. „Актуелна теренска истраживања дијаспоре. Срби у Мађарској“. *Теме 2* (2004): 847–858. Штампано.
- Сикимић, Биљана. „Тимски теренски рад Балканолошког института САНУ. Развој истраживачких циљева и метода“. Милина Ивановић Баришић (ур.). *Теренска истраживања – поетика сусрета*. Београд: Етнографски институт САНУ, 2012. 167–198. Штампано.
- Сикимић, Биљана. „Увод у етномиколошка истраживања Заплања и Тимока“. Зоја Карановић (ур.). *Гора божурова. Биљни свет у традиционалној култури Словена*. Београд: Удружење фолклориста Србије, Универзитетска библиотека „Светозар Марковић“, 2017. 211–226. Штампано.
- Сикимић, Биљана. „Антрополошко-лингвистичка истраживања Заграђа 2012“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 145–179. Штампано.
- Собољев, Андреј. „О пиротском говору у светлу најновијих истраживања“. *Пиротски зборник 21* (1995): 195–214. Штампано.
- Станковић, Ана. „Репушница: поглед са границе сећања“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018. 17–30. Штампано.
- Станојевић, Маринко. „Северно-тимочки дијалекат (прилог дијалектологији источне Србије)“. *Српски дијалектолошки зборник II* (1911). Штампано.
- Ћирковић, Светлана. „Улога истраживача у креирању корпуса конверзационих наратива“. *Филолог 11* (2015): 267–280. Штампано.
- Ћирковић, Светлана. „Увод“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018а. 7–15. Штампано.

- Ћирковић, Светлана. „Невербална комуникација у антрополошко-лингвистичком интервјуу: анализа мултимодалних транскрипата наратива на тему гајења и прераде конопље“. Светлана Ћирковић (ур.). *Тимок. Фолклористичка и лингвистичка теренска истраживања 2015–2017*. Књажевац: Народна библиотека „Његош“, Београд: Удружење фолклориста Србије, 2018б. 219–238. Штампано.
- Ћирковић, Светлана. „Угроженост животне средине као елемент антрополошко-лингвистичких интервјуа у Књажевцу и околини“. Микица Сибиновић, Владана Стојадиновић, Данијела Поповић Николић (ур.). *Књажевачки крај – потенцијали, стање и перспективе развоја*. Књажевац: Народна библиотека „Његош“, Београд: Српско географско друштво, 2019. 145–157. Штампано.
- Beal, Joan, Karen Corrigan, and Hermann Moisl (eds.). *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*. Basingstoke: Palgrave Macmillan, 2007. Printed.
- Chambers, Jack K., Peter Trudgill and Natalie Schilling-Estes (eds.). *The Handbook of Language Variation and Change*. Chichester: Wiley-Blackwell, 2003. Printed.
- Ćirković, Svetlana. “Use of the Evaluative Dative Reflexive *si* in the Timok and Lužnica Vernaculars in Eastern Serbia”. Рад саопштен на скупу *XLVIII Международная филологическая научная конференция*, 18 – 27. март 2019, Филолошки факултет Државног универзитета у Санкт Петербургу. Веб. 11. 06. 2019.
- Ćirković, Svetlana. “Who’s Afraid of the Big Bad Hemp? The growing and processing of hemp in Eastern Serbia”. Magdalena Slavkova et al. (eds.). *Between the Worlds: People, Spaces and Rituals*. Sofia: Institute of Ethnology and Folklore Studies with Ethnographic Museum, Bulgarian Academy of Sciences. In press.
- De Busser, Rik, and Randy J. LaPolla (eds.). *Language Structure and Environment: Social, Cultural, and Natural Factors*. Amsterdam: Benjamins, 2015. Printed.
- Jakubíček, Miloš, et al. “Fast Syntactic Searching in Very Large Corpora for Many Languages”. *PACLIC*, 2010. 741–47. Web. 10. 06. 2019.
- Jensen, Kim Ebensgaard. “Linguistics and the digital humanities: (computational) corpus linguistics”. *MedieKultur. Journal of media and communication research* 57 (2014): 115–134. Web. 10. 06. 2019.
- Komrsková, Zuzana, et al. “New spoken corpora of Czech: ORTOFON and DIALEKT”. *JAZYKOVEDNÝ ČASOPIS* 68/2 (2017): 219–228. Printed.

- Ljubešić, Nikola, and Filip Klubička. 2014. "{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian." Felix Bildhauer and Roland Schäfer (eds.). *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, Gothenburg, Sweden, 2014. 29–35. Web. 10. 06. 2019.
- Ljubešić, Nikola, et al. "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: ELRA, 2016. 4264–4270. Web. 10. 06. 2019.
- Makarova, Anastasia, and Teodora Vuković. *The Influence of Area and Age on the Use of Possessive Dative Clitics and Possessive Pronouns in Torlak Varieties in Serbia and Bulgaria*. In preparation.
- Miličević, Maja, i Nikola Ljubešić. „Tviterasi, tviteraši ili twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian twittes". *Slovenščina 2.0* 4/2 (2016): 156–188. Štampano.
- Mirić, Mirjana. 2019. *Areal analysis of the complementizer omission in the Prizren-Timok dialectal area of the Serbian language*. Рад саопштен на скупу *XLVIII Международная филологическая научная конференция*, 18 – 27. март 2019, Филолошки факултет Државног универзитета у Санкт Петербургу. Веб. 11. 06. 2019.
- Norrick, Neal. "Remembering and Forgetfulness in Conversational Narrative". *Discourse Process* 36/1 (2003): 47–76. Printed.
- Rychlý, Pavel. "Manatee/Bonito – a modular corpus manager". *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 2007. 65–70. Web. 10. 06. 2019.
- Savić, Svenka, i Vesna Polovina. *Razgovorni srpskohrvatski jezik*. Novi Sad: Filozofski fakultet, Institut za južnoslovenske jezike, Beograd: Studio plus, 1989. Štampano.
- Schmidt, Thomas. "Creating and Working with Spoken Language Corpora in EXMARaLDA". Lyding, V. (ed.). *Lesser Used Languages & Computer Linguistics II*. Bolzano: Eurac Research, 2009. 151–164. Printed.
- Sobolev, Andrej. *Sprachatlas Ostserbiens und Westbulgariens (III)*. Texte. Scripta Slavica Band 4. Marburg / Lahn: Biblion-Verlag, 1998. Štampano.
- Utvić, Miloš. 2013. *Izgradnja referentnog korpusa savremenog srpskog jezika*. Filološki fakultet Univerziteta u Beogradu. Doktorska disertacija. Dostupno na:

<http://nardus.mpn.gov.rs/bitstream/handle/123456789/4091/Disertacija.pdf?sequence=1&isAllowed=y> Veb. 01. 06. 2019.

Vuković, Teodora, and Maja Miličević. "Creation and some ideas for classroom use of an electronic corpus of the dialect of Bunjevci". Jelena Filipović and Julijana Vučo (eds.). *Minority Languages in Education and Language Learning: Challenges and New Perspectives*. Beograd: Filološki fakultet, 2017. 353–368. Printed.

Vuković, Teodora. "Torlak Transcription Guidelines". 6 December 2016. University of Zurich. Guidelines for project collaborators.

Vuković, Teodora. *Representing variation in a spoken corpus of an endangered dialect. The case of Torlak*. Submitted.

Vuković, Teodora. "Guidelines - Annotation". April 2019. University of Zurich. Guidelines for project collaborators.

Mirjana Mirić, Maja Miličević Petrović, Svetlana Ćirković

DIGITIZING LANGUAGE AND CULTURE THROUGH ELECTRONIC CORPORA: THE CASE OF THE TIMOK VERNACULAR

Summary

The aim of this paper is to describe the ongoing development of an electronic corpus of the Timok vernacular, a rare example of an oral dialect corpus of the Serbian language. The corpus comprises data relevant for both linguistics and studies of (traditional) culture, and as such it can help bridge the gap currently present between corpus linguistics and digital humanities. The material contained in the corpus is a result of fieldwork research conducted between 2015 and 2017, mainly within the project *Protecting the intangible culture of the Timok vernacular*. The paper outlines the phases of fieldwork research, in particular the selection of villages and participants, as well as the open-ended interview methodology applied in data collection. The steps in corpus development are presented next: transcription, annotation (part-of-speech tagging, lemmatization, normalization), and the resulting search options. In addition, an overview of previous and ongoing studies based on the collected material are provided, capturing the domains of dialectology, Balkan linguistics, socio-, areal and anthropological linguistics, as well as studies of folklore and traditional culture, with suggestions for future research in these domains.