ДИГИТАЛИЗАЦИЈА КУЛТУРНЕ И НАУЧНЕ БАШТИНЕ, УНИВЕРЗИТЕТСКИ РЕПОЗИТОРИЈУМИ И УЧЕЊЕ НА ДАЉИНУ

Тематски зборник у 4 књиге:

Књ. 1: Дигитализација културне и научне баштине
Књ. 2: Дигиталне библиотеке, дигитални репозиторијуми, дигиталне презентације
Књ. 3: Дигитални извори у друштвено-хуманистичким истраживањима
Књ. 4: Учење на даљину и интерактивна настава

Уредници:
проф. др Александра Вранеш
проф. др Љиљана Марковић
проф. др Гвен Александер

Рецензенти:
проф. др Александар Димчев
проф. др Ендрју Џ. Смит
проф. др Александар Јерков
проф. др Данијела Костадиновић
проф. др Х. Џејмс Биркс

Књ. 3:

ДИГИТАЛНИ ИЗВОРИ У ДРУШТВЕНО-ХУМАНИСТИЧКИМ ИСТРАЖИВАЊИМА

Уредници:
проф. др Александра Вранеш
проф. др Љиљана Марковић
проф. др Гвен Александер

Стана Ристић, Тања Самарџић,
Милена Јакић, Александра Марковић,
Ненад Ивановић
Институт за српски језик САНУ, Београд

УДК 004:811.163.41'374(038)
930.85:004.9

ЗНАЧАЈ ДИГИТАЛИЗАЦИЈЕ ЈЕЗИЧКИХ РЕСУРСА РЕЧНИКА СРПСКОХРВАТСКОГ КЊИЖЕВНОГ И НАРОДНОГ ЈЕЗИКА САНУ ЗА РАЗВОЈ НАУКЕ И ОЧУВАЊЕ КУЛТУРНЕ БАШТИНЕ

Сажетак

Постојећа грађа за израду *Речника савременог српскохрватског књижевног и народног језика САНУ* представља колекцију од око шест милиона листића који садрже податке на основу којих се утврђују значења, обрасци употребе и граматичке особине речи српског језика. Грађа је брижљиво прикупљана током периода од готово сто година и данас представља богат и незаменљив извор сазнања о српском језику, култури, па чак и историји. Са развојем рачунарске технологије ствара се могућност складиштења и чувања језичких ресурса Речника САНУ у дигиталном облику.

Наш рад има неколико циљева: а) истицање доприноса који би дигитализација језичких ресурса Речника САНУ дала развоју науке и очувању културне баштине; б) идентификовање конкретних поступака које би пројекат дигитализације требало да обухвати; ц) дефинисање стандарда квалитета дигитализације. Наша разматрања обухватају анализу начина на који се тренутно користи речничка грађа, испитивање могућности њене рачунарске обраде, као и процену потребних улагања. Настојимо да покажемо да овај подухват, који би омогућио лакши приступ ресурсима, као и њихово аутоматско претраживање, захтева релативно мало улагања у односу на очекивану корист не само за истраживачку заједницу која се бави проучавањем и стандардизацијом српског језика, већ и за друштво у целини.

Кључне речи: Речник савременог српскохрватског књижевног и народног језика, САНУ, дигитализација, културна баштина

1. Увод

1.1. Задатак овог рада је да представи процес дигитализације језичких ресурса *Речника српскохрватског књижевног и народног језика САНУ* (у даљем тексту: Речник САНУ) као саставни део развоја методологије савремене лексикографске обраде. Сходно томе, рад има за циљ да укаже на значај дигитализације у унапређењу и модернизацији сложеног посла израде Речника САНУ као тезаурусног речника савременог српског језика.

1.2. У складу са изложеним, рад је подељен на три тематске целине.

1.2.1. У првој целини говори се о значају Речника САНУ, као и његових језичких ресурса, за савремену српску науку и културу. Након тога, следи детаљан приказ начина на који се тренутно користи речничка грађа, као и приказ поступака које би требало предузети како би се језички ресурси овог речника адекватно складиштили у дигиталном облику.

1.2.2. У наставку рада разматрају се специфична питања дигитализације најобимнијег дела речничког корпуса: лексикографских листића (картица, фиша). У рачунарској обради овог дела грађе језичких ресурса Речника САНУ најдаље се одмакло. У истом одељку, износе се и процене потребних улагања у подухват дигитализације.

1.2.3. Коначно, у раду се експлицира улога дигитализованих језичких ресурса Речника САНУ у контексту израде *лексикографске радне станице* као рачунарске апликације која обједињује предности дигиталних корпуса са традиционалним лексикографским методама.

2. Речник САНУ и његов значај за савремену српску науку и културу

2.1. Речник САНУ представља најзначајније дело српске лингвистике и једно од најзначајнијих дела српске културе. Израђује се у Институту за српски језик САНУ, а у његовој изради данас учествују 33 стално запослена и 4 спољна сарадника.

Овај тезаурусни Речник обухвата целокупну лексику српског књижевног и народног језика у периоду од два века његовог савременог развитка. У њему се дају све језичке и употребне

карактеристике речи у виду информација различитог типа: граматичких, нормативних, терминолошких, стилских, прагматичких, културолошких, као и потврде за употребу речи у облику цитата из веома обимне грађе. Поред тога, наводи се и употреба дате речи у фразеолошким јединицама, народним пословицама и загонеткама. Лексикографско дело оваквог обима, како истиче И. Грицкат (2007: 95), пружа детаљна знања о појмовима, као и знања о самим речима као предмету науке о језику.

2.2. Грађу за израду Речника чини око шест милиона руком и писаћом машином исписаних лексикографских листића (фиша), који садрже примере употребе речи, заједно са другим подацима (о извору записа, морфолошким карактеристикама речи и сл.). Поред тога, постоје још два основна дела грађе за израду Речника САНУ: библиотечки извори и објављени томови Речника САНУ.

Тренутно се ова грађа чува и користи искључиво у папирном облику, а велики део грађе је (пре свега лексикографски листићи и библиотечки извори) у веома лошем стању. Будући да је грађа материјал од изузетног националног значаја, али и вредан извор података о култури, историји и језику нашег народа, ову грађу треба:

а. конзервирати, односно заштитити од даљег пропадања;

б. учинити доступном, тј. олакшати њену дистрибуцију; а поред тога

в. веома је важно материјал учинити лако претраживим.

Вредност ове грађе је непроцењива, а сталним коришћењем при изради Речника САНУ и другим истраживачким пројектима она се додатно оштећује. Због тога је од великог значаја добро осмислити редослед корака дигитализације и што пре отпочети сам процес, који би одговорио на три наведена задатка.

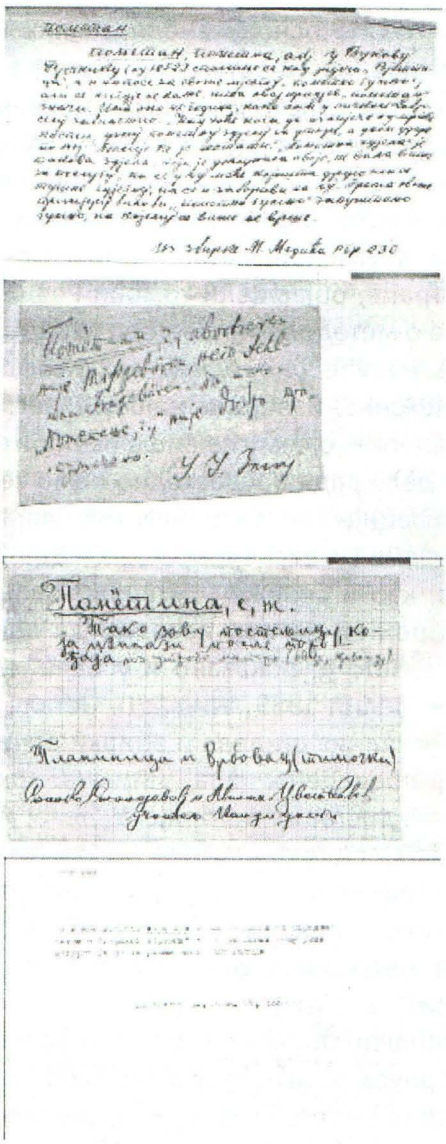
2.3. Размена искустава са институцијама које су већ приступиле процесу дигитализације материјала умногоме би помогла при дефинисању стандарда квалитета. Ипак, не треба изгубити из вида да начин дигитализације треба прилагодити специфичним потребама пројекта и захтевима при коришћењу грађе.

3. Грађа Речника САНУ – угрожено научно и културно благо

3.1. При изради речника, прикупљање грађе представља иницијалну фазу и подразумева одлуке које се тичу избора узорка језика на основу кога ће се радити речник (да ли ће се узимати само писани или и говорни језик, белетристика или стручни текстови, периодика и сл.). Након ове фазе следи одабир одредница из прикупљене грађе (одлуке које се тичу слоја лексике који ће ући у речник; тј. да ли ће се узимати архаизми, колоквијализми, дијалекатске речи, термини и сл.). Када је утврђено које ће се јединице наћи у речнику, приступа се конструисању лексикографског чланка, тј. утврђивању типова информација о свакој одредници у речнику (акценатских, граматичких, семантичких, синтаксичких и др.). Упоредо са планирањем структуре речника, утврђује се и начин на који ће бити распоређене одреднице: по азбучном реду, према творбеним језгрима, тематски и сл. (о овим фазама израде речника видети више у: Згуста 1991: 212).

3.2. Сходно наведеним фазама, пројекат израде Речника САНУ заснован је крајем 19. века, када је почело и прикупљање записа о употреби речи на лексикографским листићима (картицама, фишама). Сваки листић садржи одредничку реч, пример употребе, информацију о извору, док поједини листићи садрже и додатне податке (о значењу, акценту, морфолошким карактеристикама итд.). Изглед и садржај лексикографских листића може се видети на Слици 1:

Слика 1: Фотографије неколико лексикографских листића, јединица грађе Речника САНУ:



3.3. Лексикографски листићи представљају основни материјал на основу кога је направљен избор одредница које ће ући у речник. Поред њих, и други делови грађе имају специфичну функцију при изради Речника САНУ. На основу библиотечких извора проверава се тачност ексцерпираних података са листића, будући да су у библиотеци Института за српски језик САНУ похрањене све књиге из којих су ексцерпирани подаци. Друга важна функција библиотечког фонда јесте допуна грађе, на тај начин што се прикупљају и у изради речника користе разни терминолошки, дијалектолошки, енциклопедијски и други приручници који су објављивани током година. Са друге стране, објављени томови Речника САНУ важан су извор информација о методологији његове израде, као и инструмент на основу кога је могуће упоредити лексикографска решења за поједине типове лексике, изабрати најбоље решење и уједначити лексикографске поступке (о чему ће више речи бити у т. 7).

3.4.1. У овом делу рада превасходно ћемо се бавити приказом поступка дигитализације лексикографских листића. Подаци на листићима прикупљени су ексцерпирањем 4.369 различитих типова штампаних извора, као и бележењем појављивања речи у усменој употреби језика. Преглед извора показује да већину чине прозна белетристичка дела, што је показано и у Белићевом *Уводу у Први том Речника САНУ* (Белић 1959: VII–XXVI). Остале изворе чине дела различитих функционалних стилова и збирки речи. Од њих, највећи број извора имају преведена дела, затим збирке речи, речници, дневни и недељни листови, историјска дела, поезија, законски списи, драмска дела, енциклопедије, часописи, религијска дела, антрополошко-географска дела, уџбеници, терминолошки речници, лексикони, граматике, правописи, филозофска дела, медицинска дела, математичка, педагошка, финансијска дела, (ауто)биографије, естетике и путописи.

3.4.2. Хронолошки преглед извора показује следеће стање: централни део корпуса чине извори од 1850. до 1950. год. То је распон од 100 година употребе језика, са укупним бројем од 3.098 извора. Евидентно је да је број извора од доње границе речничке грађе до 1900. год. мањи (укупно 995); док је број извора од 1900. до 1950. године знатно већи (2.103 извора). Овај период од 50

година потврђује и стабилно стање у језичком развоју, за разлику од претходног периода, који је обележен његовим развојем на путу до стабилизације. Период од 50 година друге половине 20. века (што је и горња граница речничке грађе), репрезентован је са укупно 865 извора, што је унеколико мањи узорак од оног који представља период од 50 година друге половине 19. века. Број извора друге половине 19. и 20. века, односно доње и горње границе речничке грађе, у периоду од 50 година је приближно једнак, што указује на њену уравнотеженост у овим развојним периодима савременог српског језика.

Друга половина 20. века репрезентована је, својим највећим делом, изворима насталим између 1950. и 1960. године, када је грађа проширена знатним бројем извора са хрватске и других територија, како би се после Новосадског договора и речничком грађом утврдило језичко јединство. Затим се већи број извора запажа у периоду од 1980. до 1990. (укупно 205), у којима, осим нових дела из белетристике из свих центара српскохрватског језика, има знатан број уџбеника и приручника, речника, терминологија и енциклопедија.

Деценија која долази непосредно после публиковања првог тома Речника, између 1960. и 1970. године, одликује се мањим бројем извора (67), док тај број нагло пада у периоду 1990–2000. год. (само 36 извора), да би од 2000. до 2006. године тај број износио само 6 извора.[1]

3.4.3. Доминантни центри у развоју језика, како показују штампани извори, из којих је ексцерпирана речничка грађа, јесу Београд (са 2.114 извора) и Загреб (са 765 извора).[2] Може се рећи

1 Важно је напоменути да су извори за Речник САНУ најпре бирани према језичким (лексичко-стилским) новинама које уносе у праксу књижевне и других видова употребе српског језика. Из овог разлога, разумљиво је што се за речничку грађу у већој мери ексцерпирају они књижевни извори који култивишу поменуте новине (нпр. писци романтизма, раног реализма, постмодернизма), него извори пореклом из књижевних периода који се, у језичком смислу, ослањају на наслеђе протеклих епоха (нпр. писци позног модернизма, фолклорног реализма, и сл.).

2 Податак о најбројнијој заступљености извора у овим центрима иде у прилог ставу М. Пешикана о примарности српског и хрватског језичког израза над другим језичким изразима српскохрватског језика (в. Пешикан 1970: 26–27).

да је утицај ових центара био највећи, јер је језик писаца, независно од варијанте на којој су стварали своја уметничка дела, како показују неки извори, прилагођаван варијанти центра у коме је дело штампано (в. Белић 1959: XIII). После њих долази Нови Сад (са 211 извора), па Сарајево (са 119 извора). Ако се све ово има у виду, очигледно је да је грађа источне варијанте језика заступљена знатно већим бројем извора, који се увећава и изворима из ранијих периода развитка српског језика: Будима (76), Беча (41), Земуна (38), Пеште (4), Прага (3), Осијека (1) и Лајпцига (1).

3.4.4. У прикупљању грађе за Речник САНУ евидентан је непрекидан напор да се избалансира грађа према различитим критеријумима како би се представила лексика целокупног српског / српскохрватског књижевног и народног језика. Примена истих критеријума у функционалностилском, тематском, хронолошком, територијалном и сваком другом одабирању лексике и у књижевном и у народном језику допринела је знатној уравнотежености речничке грађе. У складу са тезама С. Новаковића о тематској, семантичкој и другој организованости лексике по „круговима употребе језика" (1893), настојало се да се у изворима обухвати целокупна реалност народног живота, рада, занимања, веровања, обичаја и друго, као и свих других општих и апстрактних појмова, заједно са припадајућим идиомима, метафорама, фигурама итд.

3.5. Сам тезаурусни Речник САНУ, заснован на оваквој грађи, представља значајан извор како за изучавање српског језика тако и за изучавање културне и историјске баштине његових носилаца, јер детаљно представља полиглосију свих језичких идиома: језичких варијанти (источне и западне), стандардних изговора (екавског и ијекавског), језичких израза (српског, хрватског, босанскохерцеговачког и црногорског), као и сву шароликост њихових дијалекатских основица.

4. Рад са лексикографским листићима при изради Речника САНУ

4.1. Лексикографски листићи за израду речника подељени су на секције, које се као делови једног тома обрађују, редигују и технички

сређују у више фаза, почев од основне обраде, преко (помоћне) редакције, суредакције и техничке редакције. Свака секција садржи од 2.500 до 4.500 листића основне грађе, накнадно ексцерпиране грађе, затим грађе контролних одредница из већ урађених секција или томова; а посебно се уз сваку секцију прилаже додатна, тзв. „Матичина грађа"[3] и ономастика. Због овакве структурираности секција обрађивач је дужан да, пре почетка рада, уазбучи све листиће из своје секције и да списку одредница из грађе дода списак одредница из обавезних приручника, тј. других релевантних речника, енциклопедија и разних терминолошких речника. Уколико се у задуженој секцији пронађе грађа из неке друге секције, она се предаје техничком уреднику који је задужен за техничку организацију рада на Речнику.

4.2. После уазбучавања и сређивања грађе, обрађивач по азбучном реду издваја и успоставља одредничку реч у њеном основном облику: глаголе у инфинитиву, именице и именске речи у номинативу са ознаком рода и облика за родове; непроменљиве речи према врстама, и сл.

4.3. Све листиће који показују употребу дате речи обрађивач је дужан да пажљиво прегледа, како би одредио сва могућа значења и друге лингвистичке, прагматичке и културолошке карактеристике дате речи, на основу чега утврђује класу јединица којој дата реч припада, а самим тим и модел по коме се та класа речи обрађује. Поред објављених томова речника, обрађивач користи интерни приручник *Упутства за израду Речника САНУ,* који служи за техничко и лексикографско уједначавање обрада; као и релевантну лексикографску и лингвистичку литературу ради што квалитетније и потпуније обраде дате речи.

4.4. Код полисемичних речи, које су обично потврђене великим бројем примера (неке одреднице имају и по 500 и више примера), обрађивач, на основу властитог лексикографског искуства које

3 После завршетка шестотомног *Речника српскохрватскога књижевног језика* Матице српске (1976), ксероксиране копије лексикографских листића те речничке грађе прикључене су основној грађи Речника САНУ, пре свега због тога што је до тада у њој била доста заступљена грађа из хрватских извора.

обухвата и знање о претходним обрадама, грубо одабира листиће односно примере којима ће илустровати употребу дате речи и на задовољавајући начин, користећи унапред прописана правила о употреби примера, дати све потребне информације о тој речи. Тек после таквог грубог избора листића из расположиве грађе за сваку реч, обрађивачу предстоји обрада листића (фиша): проверавање тачности навођења одредничке речи, ваљаност наведеног примера, идентификовање извора, навођење године, по потреби и места издања штампаног извора, уношење библиотечке сигнатуре извора (ради проналажења истог у библиотеци), како би се пример проверио или уобличавањем прилагодио лексикографском чланку и др. За листиће који представљају збирке речи обавезно је идентификовање сакупљача, као и навођење места односно ширег подручја у коме је реч засведочена.

4.5. Тек после завршеног рада на изабраним листићима, лексикограф приступа стручном раду, тј. граматичкој обради одредничке речи, утврђивању етимологије за речи страног порекла, навођењу квалификативних ознака, маркера којима се одређује домен употребе дате речи, детаљној семантичкој анализи примера ради издвајања значења, хијерархијском устројавању значења и њиховом што прецизнијем дефинисању, уз издвајање израза и пословица, који се обрађују у посебном делу речничког чланка. Затим следи компјутерски унос текста уз примену прописаних техничких поступака (тип и величина слова, употреба одређених знакова интерпункције, бројчано и словно издвајање значења, правилна употреба техничких скраћеница и др.). Завршена верзија прве фазе израђеног текста у електронској и штампаној форми, заједно са листићима обрађене секције, предаје се техничком уреднику.

4.6. Поменути поступци понављају се и у наредним фазама израде Речника САНУ (помоћна редакција, редакција, суредакција) и проширују даљом применом методолошких решења везаних за корекцију обраде и стандардизацију обрађених одредница у општем лексичком фонду.

5. Предности дигитализације лексикографских листића

5.1. Процес рада са листићима, у коме се утврђује и проверава тачност свих наведених података о датој речи, одузима велики део времена не само основним обрађивачима, него и лексикографима у вишим фазама израде (помоћним и основним редакторима и суредакторима). Наиме, будући да Речник САНУ служи, поред осталог, и као поуздан референтни извор на који се позивају његови корисници, подаци који су наведени у њему морају бити тачни. Из овог разлога, процедура израде Речника САНУ обухвата више нивоа провере унетог текста.

5.2. Резултат подухвата преношења грађе из папирне у електронску верзију требало би, дакле, да буде база података у којој би се нашле све информације из грађе, верно и дословно пренете. На тај начин добили бисмо поуздан извор података који лексикографи не би морали даље да проверавају и на који би се могли директно позивати и истраживачи из осталих области. Сам поступак провере података приликом дигитализације усложњава процес и чини га захтевнијим, али овакав приступ је свакако неопходан.

5.3. Са друге стране, оптималним коришћењем рачунарске технологије неки поступци провере података могли би се аутоматизовати. Сравњивање одредничких речи, на пример, са обавезним приручницима, тј. са другим релевантним речницима, енциклопедијама и разним терминолошким речницима, такође би се могло обавити аутоматски. И остали поступци техничке обраде листића као што су уазбучавање, идентификација извора и уношење библиотечких сигнатура, идентификација сакупљача и сл, могли би се аутоматизовати.

5.4. Укључивање поступака провере података и техничке обраде листића у процес дигитализације, уз аутоматизацију појединих поступака, омогућило би системски приступ техничкој обради грађе, што би довело до конзистентније обраде и до смањења броја грешака, чиме би се задовољио критеријум акрибичности.

5.5. Како би дигитализација грађе резултирала функционалним и корисним ресурсом, било би неопходно имати у виду још један

аспекат рада. Наиме, база података морала би да буде осмишљена тако да се може накнадно допуњавати новим подацима и новим речима, у складу са савременим лексикографским захтевима. Евентуална накнадна анотација постојећих листића и подаци који би се уносили, детаљније се разматрају у раду С. Ристић и Н. Ивановића (2011: 529–553).

6. Финансијска и временска процена улагања

6.1. Избор методе дигитализације текстуалне грађе прилагођава се конкретним потребама и циљевима дигитализације. Уколико се жели конзервирати стара и вредна грађа приступа се микрофилмовању, као једином универзално признатом начину архивирања грађе. Повећање доступности материјала могуће је постићи скенирањем. Ипак, у оба поменута случаја претраживост самог материјала је веома ограничена. Претрага материјала могућа је искључиво уколико се материјал претвори у текстуални документ прекуцавањем текста или оптичким препознавањем карактера након скенирања материјала. Напредна претрага пак могућа је једино ако се тако припремљен материјал додатно анотира.

6.2. На основу искустава институција које су већ приступиле дигитализацији текстуалне грађе сазнајемо да покушаји оптичког препознавања карактера штампаног материјала нису довели до задовољавајућег процента тачности. Тачност прочитаног садржаја зависи од неколико фактора: квалитета штампе, нијансе подлоге, физичке очуваности странице (искрзане или замрљане странице представљају проблем за ове програме) итд. Стога књиге морају бити у добром стању како би се могле аутоматски претворити у текст. Досадашња искуства са оптичким препознавањем говоре да чак и најбољи програми не доносе потпуну тачност. Проценат грешке често чак и након отклањања систематских грешака износи 10% на нивоу речи[4], што би значило да на једној ауторској страни на задати упит не можемо добити резултат за 180 речи. Овакав проценат чини нам се неприхватљивим. Узевши притом у обзир да библиотечку грађу

4 Овај податак добијен је на основу скенирања добро очуваних књига.

Речника САНУ чине оригиналне књиге из којих је вршена ексцерпција примера подвлачењем, као и да су неке од књига у физички лошем стању, понеке писане помоћу старе графије, а у већини приручника који се користе означени су акценти и сл. постаје јасно да ово није решење које би задовољило потребе лексикографа на пројекту израде Речника САНУ, барем не за читав библиотечки фонд. При дигитализацији рукописне грађе (а управо овом типу припадају лексикографски листићи који чине основу лексикографске грађе за Речник САНУ) проценат тачности препознатих карактера био би далеко нижи, тј. грешка далеко већа. Поготово када се има у виду да на преко 3 милиона листића постоји више десетина, можда и неколико стотина различитих рукописа. Оваква ситуација сугерише да се претраживи материјал не може добити скенирањем, након чега би следило оптичко препознавање.

6.3. Далеко већи проценат тачности пак добија се дактилографским прекуцавањем материјала. При оваквом начину дигитализације најчешће се користи принцип дуплог уноса (тзв. Double king), тј. унос материјала од стране два независна дактилографа. Ипак, у случају грађе Речника САНУ дупли унос не би био потребан, будући да се сваки пример проверава, тј. упоређује са оригиналним извором, како би се задовољио критеријум потпуне акрибичности. Поред тога, тачност материјала добијеног на овај начин у просеку износи 99,99%, што би значило да на једној ауторској страни у просеку постоји две грешке, и то рачунато на нивоу карактера. Уколико се изабере овакво решење, остаје питање који формат треба изабрати за прекуцавање материјала.[5]

6.4. У оквиру припреме послова дигитализације грађе за РСАНУ изведена су и три пилот-огледа. Ови огледи тицали су се

5 Могућа решења су: прекуцавање у текст едиторима, типа WORD-а (што би било једноставно али не и најфункционалније решење), прекуцавање у табеларном софтверу, типа EXCEL-а (што би било функционално, али не и ергономско решење за корисника апликације), и прекуцавање у базу података, што би задовољило и захтев једноставности уноса са корисничке тачке гледишта, али и захтев функционалности и искористивости унетог материјала. Апликација за унос профилисала би се на основу конкретних захтева и потреба овог пројекта, у сталној сарадњи лексикографа и програмера.

дигитализације листића, дакле рукописне грађе, а основни задатак био је процена времена и потребних улагања.[6]

6.5. Оглед бр. 1 – Скенирање листића

6.5.1. Први оглед спроведен је у склопу разматрања похрањивања фотографија листића у дигиталном облику. Ово се сматрало „брзим" решењем, чија би претраживост у потпуности зависила од додатног анотирања базе, будући да се рукописна грађа не би могла пропустити кроз процес оптичког препознавања карактера, тј. фотографије текста нису претраживи текст. У овом огледу учествовала су два цивилна војника са Математичког факултета. Они су скенирали секцију П-67, тј. кутију речничке грађе сачињене од 4.152 лексикографска листића. Од 4.152 листића 47 листића није се уопште могло прочитати са слике, тј. отисак на слици био је блед (примери скенираних листића дати су на сл. 1). Имајући у виду да учесници овог огледа нису били посебно обучени за напредна подешавања у софтверу за скенирање, сматрамо да би проценат читљивости могао бити и бољи, али на рачун брзине посла, будући да свако додатно подешавање, које зависи од конкретног листића (боја позадине, искрзаност листића, јачина трага оловке којом је листић писан, однос боје позадине и слова и сл.) захтева одређено време. Ипак, да поновимо, материјал добијен на овај начин није претражив без додатног анотирања (додатне ознаке могле би нпр. бити тип одреднице која је обрађена на листићу, извор, година издања, функционални стил и сл.).

6.5.2. Овај посао урађен је за 4 човек-дана, где је просечно трајање једног радног дана било 6 сати, а листићи су на скенер слагани у групама од по најмање 4 листића (на скенеру формата А4) а највише 8 листића (на скенеру формата А3). У табели 1, приказана је анализа овог огледа.

---

6 Као што је већ речено, постоје три основна дела грађе која би се дигитализовала (лексикографски листићи, односно фише, приучници и технички документи, као и објављени томови Речника САНУ), те се стога читав процес може одвијати у фазама, уз детаљну анализу који би редослед поступака био најцелисходнији. Дакако, неки процеси се могу одвијати и паралелно. Исто тако, није потребно чекати да се заврши унос целокупног материјала. Рад може почети у тренутку када је унесено једно слово или чак и само неколико секција.

Табела 1. Анализа временских, материјалних и финансијских улагања на основу огледа 1, рађеног на опреми доступној у Институту за српски језик:

оглед бр. 1 скенирање листића	април, 2010
број листића у узорку	4.152
процењен број преосталих листића од П до Ш	3.000.000
број непрочитаних листића	47
% непрочитаног материјала	1.13
приближна пројекција времена у годинама (10 сарадника по 4 сата дневно)	2
тип улагања	опрема, обука, цена сата[7]
додатно време	обрада материјала (раздвајање слике на листиће, тј. кроповање и сл.)

6.5.3. Као што се из табеле види поред времена потребног оператеру да скенира листиће, додатна улагања била би: обука оператера за рад са машином, обука оператера за напредна подешавања, набавка опреме, накнадна обрада материјала, анотирање добијеног материјала и сл.[8]

7 Имајући у виду цену радног сата оператера за машином, као и време потребно за овај део посла дигитализације, лако је израчунати потребна финансијска улагања, с тим што треба урачунати и цену опреме, као и обуку особља.

8 Брзина напредних скенера формата А4 износи до 100 листова у минути, али су процене рађене за папир А4 формата гранулације 80g. Остаје да се емпиријски утврди да ли би такав апарат могао да скенира листиће формата 11x17 цм или 10x14 цм, различите гранулације (картона на које су налепљени листићи од папира лошег квалитета, старих, делимично иструлилих и искрзаних листића), истом брзином, са задовољавајућим квалитетом читљивости слике, а да их притом не оштети. Мало је вероватно да би сви ови захтеви били испуњени, али уколико би било тако 1 оператер би у идеалном случају за 100 дана, радом од 5 сати дневно скенирао преосталих 3 милиона листића, на машини која скенира 100 листића у минути. Ипак, треба запамтити да таква грађа и даље не би била претражива.

6.6. Оглед бр. 2 – Дактилографско прекуцавање листића у текст едитору

6.6.1. Овај оглед за циљ је имао процену улагања и тачности добијеног материјала, који би био претражив без додатног анотирања. Показало се да је овај тип дигитализације погоднији од претходно приказаног, будући да је број потребних радних сати једнак броју сати за скенирање, а нема као у претходном случају потребе за другим пословима као што су припремне радње обуке и накнадне обраде материјала. Тачност и претраживост овако добијеног материјала је готово потпуна. У табели се могу видети основни подаци везани за овај оглед.

Табела 2. Анализа временских, материјалних и финансијских улагања на основу огледа 2:

оглед 2: прекуцавање листића	2007
број листића у узорку	384
процењен број преосталих листића од П до Ш	3.000.000
број непрочитаних листића	0
број непрочитаних знакова	344
% непрочитаног материјала	0.8
приближна пројекција времена у годинама (10 сарадника по 4h дневно)	2
улагања	цена сата

6.6.2. Имајући у виду горе наведене податке, десет дактилографа цео материјал би унело за око две године (ако би радили четири сата дневно). Поређењем ова два огледа видимо да би и временска и новчана улагања у дигитализацију путем скенирања била знатно већа, док би искористивост добијеног материјала била драстично мања.

6.7. Оглед бр. 3 – Прекуцавање листића у посебном формату компатибилном са базом података, уз лакше лексикографске послове везане за обраду листића.

6.7.1. Циљ овог огледа био је спецификација захтева за софтвер помоћу кога би се дигитализовали и обрадили листићи.

6.7.2. У овом огледу учествовао је волонтер, студент постдипломских студија Филолошког факултета. Он је у унапред дефинисану симулацију базе уносио податке са листића. Поред прекуцавања листића (што је посао дактилографа), сарадник је обављао још два додатна посла. Будући да је на листићу дата нестандардна скраћеница извора, он је на основу *Скраћеница* (интерног документа у коме се налази списак извора за грађу Речника САНУ са стандардним шифрама) придодељивао стандардну скраћеницу, а потом и сигнатуру извора у библиотеци Института за српски језик САНУ где се чувају све књиге из којих је вршена ексцерпција. Брзина његовог уноса била је далеко мања од брзине у претходном огледу (1:10), што објашњавамо као последицу три фактора:

а. лексикографски приправници, као ни искусни лексикографи нису специјализовани за брз унос текстуалних података;

б. унос података је био комплекснији, будући да се ради о изгледу сличном EXCEL табели, за разлику од прекуцавања у уобичајеним текст едиторима (прављење специјализоване апликације елиминисало би фактор комплексности уноса); и

в. у овом огледу постојала су два додатна посла, која дактилограф у претходном огледу није обављао.

6.8. На основу обављених пилот-огледа закључујемо да је, имајући у виду потребе коришћења материјала на изради Речника САНУ, најоптималнији начин дигитализације лексикографских листића прекуцавање. На страни оваквог избора нису само временска динамика и финансијска улагања, већ и квалитет добијеног материјала изражен и према проценту тачности и према претраживости текста.

7. Примарни (лексикографски) циљеви дигитализације језичких ресурса Речника САНУ

7.1. У дефинисању примарних циљева дигитализације лексикографских листића као најважнијег језичког ресурса Речника САНУ, као и осталих његових ресурса (библиотеке, објављених томова Речника, приручника, допунске грађе...) неопходно је имати у виду *стручне,* као и *техничке потребе* лексикографског тима ангажованог на изради самог Речника.

7.1.1. *Стручне потребе* за дигитализацијом засноване су на идеји да је у раду са рачунарским базама језичких података могуће проширити знања о лексикографској обради и о представљању језичких појава у Речнику САНУ. Увидом у електронски претражив корпус уређених језичких информација, који би био обликован према њиховим захтевима, лексикографима би се омогућило да потпуније сагледају језичку грађу којом располажу, као и да се у обради позабаве лингвистичким подацима који до сада нису узимани у обзир, попут: статистичких информација о граматичким, семантичким и стилским обрасцима у којима се нека реч специфично јавља; или података о фреквенцији употребе појединих значења речи, њихових морфолошких облика и сл. (према: Мајс 1996: 102–103).

Другим речима, насупрот *интерпретацији језичких информација* као поступку парафразирања значења и контекста употребе лексема у традиционалној лексикографији; дигитализација речничких ресурса подразумева, у крајњој линији, *генерисање језичких информација*, као поступак дефинисања лексема на основу велике количине аутоматски уређених података. Као предности оваквог начина рада обично се истичу: могућност да се речнички чланак уреди на основу већег броја системских дефиниција; могућност да се језички концепти, у речничкој обради, повежу са одређеним значењима; као и могућност да се, међу елементима лексикона, сигурније утврде релације на скалама синонимија : антонимија и хиперонимија : хипонимија (према: Ђинг и сар. 1997), што за крајњи резултат има садржајнији речнички текст који се нуди кориснику.

7.1.2. *Техничке потребе* дигитализације условљене су планом за убрзање посла на изради Речника САНУ, уз истовремено очување

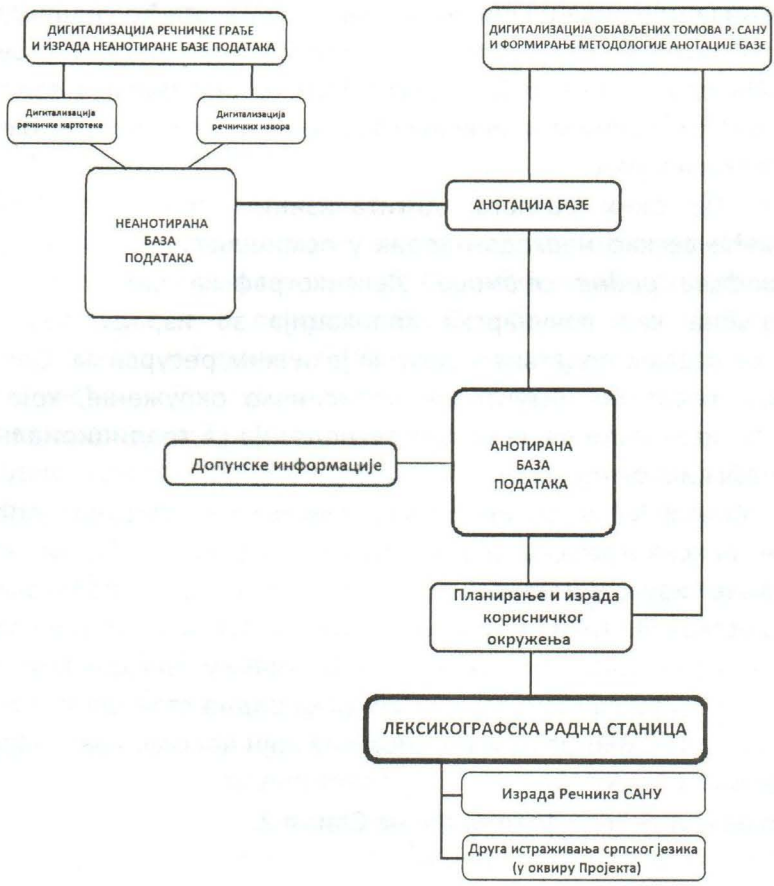
високих стандарда његове израде. Стога се од језичких ресурса Речника САНУ у дигиталном облику очекује и да послуже као основ боље организације рада на пројекту, најпре тиме што ће унапредити могућности интеракције међу сарадницима који раде на овом Речнику. На овај начин, биће могуће јасније дефинисати задатке индивидуалних учесника у лексикографском тиму и координисати њихов заједнички рад.

7.1.3. Из ових разлога, дигитализација грађе за Речник САНУ намеће се као неоходан корак у осмишљавању и креирању *лексикографске радне станице.* Лексикографска радна станица је замишљена као рачунарска апликација за израду речника, повезана са базама података и другим језичким ресурсима. Оваква апликација имала би интуитивно корисничко окружење, које би објединило предности рачунарских технологија са традиционалним лексикографским методама.

7.2. Узимајући у обзир велику сложеност описаног посла, креирање лексикографске радне станице одвијало би се кроз следеће фазе: израда (неанотиране) базе података од података са лексикографских листића; дефинисање методологије анотације базе; осавремењивање базе додатним лингвистичким информацијама. На крају овог процеса стоји лексикографска радна станица као мост између знања лексикографа и потенцијала који поседује рачунарски уређена језичка база.

Приказ изложених фаза дат је на Слици 2.

Слика 2: Фазе дигитализације језичких ресурса Речника САНУ:



У наставку излагања дајемо опис сваке фазе, после чега ће уследити закључак који резимира наводе изложене у раду.

7.3. Дигитализација речничке грађе и израда неанотиране базе података

У првој фази овако осмишљеног пројекта спроводи се дигитализација речничке грађе. Грађа за израду Речника САНУ обухвата картотеку лексикографских листића са примерима, као и библиотечке изворе који служе као њихова потврда (уп. т. 3.3.).

7.4. Дигитализација листића

7.4.1. Картотека Речника САНУ, на основу које се он израђује, састоји се од лексикографских листића са примерима. Примери су бирани за сваку реч (одредницу) понаособ.

На већини листића постоје следећи типови информација:

а. *циљна реч* која је предмет обраде, а у речнику је наведена као одредница;

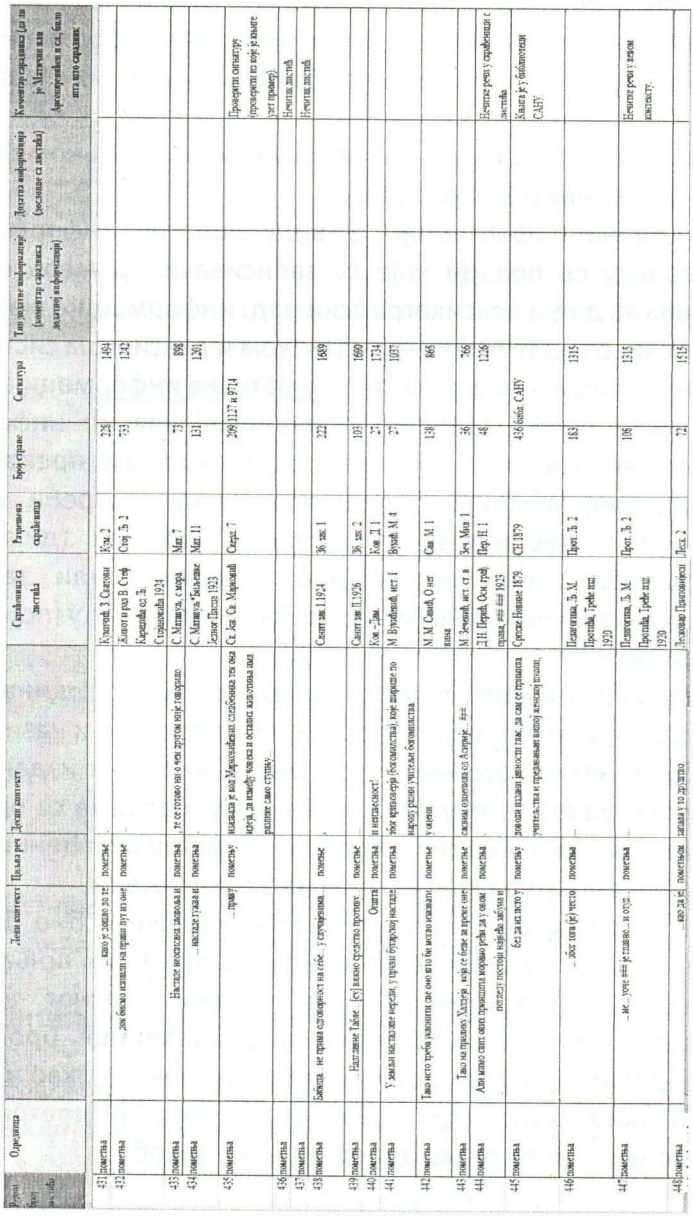
б. *конкордансија,* тј. контекст са подвученом циљном речи, који обично садржи леву и десну страну;

в. *додатна информација.* У виду додатне информације на листићу јављају се подаци које су записивачи примера сматрали релевантним за даљи лексикографски рад: информације о значењу и употреби речи, о месту речи у граматичком и лексичком систему и сл. Тако се у овим пољима могу јавити: граматичке информације (подаци о припадности речи одређеној граматичкој класи); информације о фразеолошкој употреби лексеме (од колокација преко идиома до израза); информације о терминолошкој употреби лексеме; информације о нестандардној употреби лексеме (дијалекатски облици; односно лексеме као део пословице или загонетке); информације о значењу лексеме у одређеном контексту (појединачна значења), и др.

г. *белешка о извору примера* (нпр. *за књиге:* година и место издања, број стране у књизи; *за збирке речи:* порекло и назив збирке; *за периодику:* назив часописа, број, месец и година издања, аутор чланка, број стране). У току рада на Речнику, белешка се претвара у стандардизовану скраћеницу на основу регистра скраћеница, а тада јој се додаје и сигнатура коју извор има у библиотеци.

7.4.2. Садржај овако структуриране картотеке био би пренет у текстуалну базу. База би се састојала од посебних поља за унос: одреднице, левог контекста, циљне речи, десног контекста, скраћенице са листића, стандардизоване скраћенице, броја стране, сигнатуре, додатне информације (са образложењем), као и пратећег коментара онога ко уноси грађу. Базу би било могуће претраживати у целини, као и у оквиру сваке одреднице понаособ.

Слика 3: Нацрт дигитализоване речничке картотеке. *Лево:* поља за унос примера. *У средини:* поља за унос извора. *Десно:* поља за додатне информације и коментаре.



7.5. Дигитализација библиотечких извора

Дигитализација библиотечких извора, који служе као потврда језичким информацијама на листићима, има двојаку функцију. Са једне стране, на основу рачунарски претраживог текста извора могуће је брзо проверити тачност уноса у картотеку. Са друге, могућност позивања на текст извора у току обраде омогућила би лексикографима да сагледају шири контекст примера (у распону од неколико реченица, пасуса, па до стране). Овим се задовољава једна од основних нужности у лексикографској обради: провера и евентуална допуна контекста у коме се јавља пример који се уноси у Речник.

7.6. Дигитализација објављених томова Речника САНУ

7.6.1. Паралелно са дигитализацијом листића и библиотечких извора, одвијала би се и дигитализација објављених томова Речника САНУ, који такође представљају значајан језички ресурс. Израдом дигитализоване верзије Речника, лексикографски тим би стекао контролни корпус за проверу свих релевантних информација у раду, и то у областима искористивости грађе из корпуса, расподеле грађе према жанровским, социолингвистичким и др. критеријумима.

7.6.2. Са друге стране, научним увидима у методолошке поступке којима се у Речнику САНУ обрађује речничка грађа, лексикографи би били у могућности да израде методологију даље анотације језичке грађе у претходно формираној бази.[9]

7.6.3. После превођења Речника САНУ у машински читљив текст, извршила би се провера метода обраде грађе у том речнику. Поменута фаза би подразумевала обиман лингвистички рад, који

9 Анотација је процес при коме се лексичкој грађи у корпусу, машински (аутоматски), ручно, или комбинованом методом (аутоматска анотација уз проверу) додају различити типови лингвистичких и других информација, релевантних за њену обраду и представљање у речничком чланку. Анализом, односно синтезом и рачунарским претраживањем анотираних информација у корпусу долази се до увида у релевантне лингвистичке појаве, значајне за систематску обраду лексике, као и за одређивање типова информација (нпр. за типско дефинисање значења, или за организацију типова информација у речничком чланку).

би обухватио уочавање и анализу веза између лексикографских поступака и различитих нивоа анализе језика. Овако стечено знање (како лексикографско, тако и програмерско) било би искоришћено при даљој изради базе података.

Прелиминарни увиди у објављене томове Речника САНУ указују на то да би се грађа у бази података могла анотирати следећим типовима информација који су у њима доступни:

а. *подаци о припадности речи одређеној лексичко-граматичкој класи или подкласи.* Поред груписања одредница према систему граматичких категорија, рад на речнику био би олакшан и тиме што би се скренула пажња на групе атипичних случајева који се регуларно јављају у лексикографској обради. Ту спадају нпр. лексеме које немају целу граматичку парадигму (јављају се само у множини, само у одређеном виду и сл.); или лексеме које егзистирају у више комплементарних граматичких категорија (глаголи и свршеног и несвршеног вида; именице и мушког и женског рода и сл.);

б. *подаци о синтаксичким функцијама речи у тексту.* Упућивањем на функцију одређене лексичке јединице као на тип релације према другим јединицама лексичког система у контексту реченичне целине, скренула би се се пажња на специфичне семантичке парадигме у којима се реализује. Овај тип анотације обухватао би дефинисање типова конструкција у којима се лексичка јединица налази, опис њених рекцијских допуна и сл.;

в. *подаци о фреквенцији речи у лексичком фонду.* Информацијама о припадности речи корпусу од 5.000, 3.000 или 2.000 најфреквентнијих речи у језику обрађивачима би се скренула пажња на то да је њену обраду потребно обогатити додатним информацијама, које не треба стављати код речи мање фреквенције (нпр. граматичко-акценатска парадигма, изузеци различите врсте, идиоматске конструкције, прагматички контексти употребе и сл.);

г. *време настанка речи и њено место у новијем развитку савременог српског језика.* Организација ове врсте података могла би се вршити по годинама, деценијама, или по кључним тачкама историјског развитка српског језика (нпр.: пре и после 1847; пре и после 1914; пре и после 1944). На тај начин би се постигла уравнотежена

заступљеност извора из различитих епоха новијег развитка српског језика у грађи;

д. *припадност речи карактеристичном жанру.* Додавањем података о жанру коме реч (идиом, фраза, конструкција) припадају (нпр. етнографија, социологија, политика, астрономија, географија и сл.) утврдили би се регистри у чијим доменима се реализују поједина (терминолошка) значења. Коришћењем ових информација у лексикографској обради створила би се добра основа за обележавање припадности речи одређеној стилској или комуникативној сфери.

7.7. Допунске информације

Сагласно једном од основних начела организације рада на Речнику САНУ: да се на систематски начин обради и представи лексичка грађа која, својом репрезентативношћу, супституише језичку целину – у базу би била унета и нова грађа, која на хронолошком, дијалекатском, територијалном, функционалностилском, тематском и семантичком плану допуњава постојећу. Избор извора за нову грађу био би поверен посебном редакционом одбору, док би избор самих примера према лингвистичким критеријумима обављао за то одређен тим лексикографа.

7.7.1. *Књижевни језик.* Требало би настојати да се новоформирана лексичка база речника равномерно обогаћује грађом српских екавских и ијекавских писаца друге половине 20. и почетка овог века, као и одговарајућом грађом из разговорног језика. При томе би постојеће разлике на фонетском и лексичком нивоу, као и двојство изговора и писма, требало третирати као особености јединственог српског језика, које у лексикографском опису није потребно посебно коментарисати.

7.7.2. *(Раз)говорни језик.* Посебно место у новој грађи заузимао би (раз)говорни језик (језик конверзације), као и они регистри који га фиксирају (као нпр. језик интервјуа, језик Интернет-комуникације и сл.).

7.7.3. *Грађа из научних студија.* Поред корпуса књижевног и (раз)говорног језика, грађа за Речник САНУ могла би се допуњавати и из научних студија које су посвећене појединим питањима развитка

српског језика (нпр. студије глаголских парадигми; творбено-семантичке студије, културолошки оријентисане студије лексике и сл.). Поменуте студије помогле би да се расветле тематска питања у лексикографској обради, а њихови индекси послужили би као допуна списку одредница Речника САНУ.

7.8. Лексикографска радна станица

Све описане фазе рада посматрамо као етапе у изради *лексикографске радне станице.* Лексикографска радна станица је рачунарска апликација повезана са анотираном базом података и другим потребним ресурсима у дигиталном облику.

7.8.1. Један од основних захтева везан је за њено *корисничко окружење.* Она треба да пружи лексикографу уобичајени (познати) радни амбијент, надграђен предностима које пружају рачунарске технологије.

Из овог разлога, корисничко окружење лексикографске станице требало би да, у што већој мери, очува постојећи процес израде Речника САНУ, на тај начин што ће симулирати радно окружење лексикографа, а истовремено му омогућити да користи напредне могућности умрежавања и доступности велике количине обрађених информација.

7.8.2. Истовремено, лексикографска радна станица би требало да садржи и резултате лексикографског рада у свакој фази обраде речничког текста, од основне обраде до техничке редакције.

Очекује се да ће овако осмишљена лексикографска радна станица унапредити рад на Речнику САНУ у областима:

а. ефикаснијег сагледавања статистичких информација о граматичким, семантичким и стилским обрасцима у којима се нека реч специфично јавља (типова граматичких конструкција, семантичких категорија и поткатегорија, функционалних стилова); односно сагледавања специфичности употребе речи у одређеним типовима текстова, или у одређеним говорним ситуацијама;

б. коришћења информација о фреквенцији употребе: речи у лексичком фонду, појединих значења речи, или појединих флективних наставака речи, које помажу лексикографима да утврде

редослед значења у речничком чланку, као и да сигурније разлуче хомонимију од полисемије;

в. могућности сталног освежавања корпуса новим речима, њиховим новим значењима и колокацијама, што лексикографима омогућава да ефикасније прате и проучавају процесе деривације, лексичког слагања, семантичког развоја речи, као и процес формирања нових идиома у лексичком систему (у оквиру овог, и могућност комуникације са корисницима и уважавања специфичних корисничких захтева);

г. шире заступљености и употребе примера из говорног језика, доследно пренетих или прилагођених лексикографовим потребама, у интервенцијама на допуни и осавремењивању значења речи.

7.8.3. Поред убрзања израде Речника САНУ и побољшавања квалитета лексикографске обраде, дигитализација његових језичких ресурса има значаја и у очувању културне баштине и у унапређењу лингвистичких истраживања српског језика.

8. Закључак

Све што је изложено у раду може се резимирати у следећим тезама:

8.1. Грађа за речник САНУ, која је, својим главним делом, сачињена од записа на листићима, представља културно богатство (заштићено законом као споменик културе), које је изложено ризику пропадања. Дигитализација би омогућила очување ових података, после чега би било могуће микрофилмовањем конзервирати саме листиће јер се њима не би више манипулисало.

8.2. Најбољи метод за дигитализацију грађе је њено прекуцавање од стране професионалних дактилографа. Грађа у дигиталном облику била би поуздана и прегледна, доступна широком кругу истраживача.

8.3. Дигитализација језичких ресурса Речника САНУ омогућила би не само њихову заштиту и очување, већ и бољу прегледност, као и лакшу и ефикаснију доступност широј научној заједници. Ипак, примарни циљеви дигитализације језичких ресурса Речника САНУ остају у домену развоја лексикографије, где је овај процес пресудан за формирање *лексикографске радне станице*, која обједињује

традиционалан лексикографски рад са предностима рачунарских технологија у раду са базама језичких података.

8.4. Дигитализација језичких ресурса је изузетно значајан и сложен подухват, који би српска научна и културна заједница морала да подржи и реализује.

ЛИТЕРАТУРА:

Белић 1959: Белић А., *Увод*, Речник српскохрватског књижевног и народног језика САНУ, VII–XXVI.

Грицкат 1993: Грицкат-Радуловић И., *Стогодишњица лексикографског рада при Српској академији наука и уметности*, Сто година лексикографског рада у САНУ, Београд, 5–13.

Грицкат 2007: Грицкат-Радуловић И., Наука о језику у делатности Академије, *Шездесет година Института за српски језик САНУ, зборник радова,* књ. I. Београд: Институт за српски језик САНУ. 19–106.

Ђинг и сар. 1997: Hongyan Jing, Kathleen McKeown, Rebecca Passonneau, "Building a rich large scale lexical base for generation", Department of Computer Science, Columbia University, New York, http://academiccommons.columbia.edu/catalog/ac:110195.

Згуста 1991: Згуста Л. *Приручник лексикографије*, Сарајево; Свјетлост, Завод за уџбенике и наставна средства.

Ивановић 2007: Ивановић Н., „Принципи формирања и организације корпуса Речника српскохрватског књижевног и народног језика САНУ (у периоду од 1853. до 1953. године)", *Шездесет година Института за српски језик САНУ, зборник радова,* књ. II. Београд: Институт за српски језик САНУ. 53–78.

Јошић 2008: Јошић Н., Корпус Речника САНУ: ријечи из народних говора и њихова регионална заступљеност", *Српски језик у (кон) тексту, зборник радова,* књ. 1, Крагујевац: ФИЛУМ. 437–447.

Meijs 1996: Meijs W."Linguistic Corpora and Lexicography", Annual Review of Applied Linguistics, vol. XVI, Cambridge: Cambridge University Press. 99–114.

Михаиловић 2007: Михаиловић, Д., *Мајсторско писмо*, Београд.

Новаковић 1893: Новаковић С., *Предлог Српској краљевској академији учињен 5. априла 1893, да се отпочне купљење грађе за академијски Српски Речник, и да се за тај посао установи у Академији Лексикографски Одсек,* посебно издање (прештампано из листа „Јавор"), Земун: Штампарија Јове Карамата.

Пешикан 1970: Пешикан М., *Наш књижевни језик на сто година послије Вука*, Београд.

Позив и Упутство 1899: *Позив и упутство за купљење речи по народу за речник Српске краљевске академије*, Штампано у штампарији Краљевине Србије, Београд.

Ристић 2006: Ристић, С., *Раслојеност лексике српског језика и лексичка норма*, Институт за српски језик САНУ, Монографије 3, Београд.

Ристић 2007: Ристић С., „Прва лексикографска школа у Институту за српски језик САНУ", *Шездесет година Института за српски језик САНУ, зборник радова,* књ. I. Београд: Институт за српски језик САНУ. 131–149.

Ристић 2008: Ристић С., „Корпус Речника српскохрватског књижевног и народног језика САНУ са становишта репрезентативности савременог српског језика", *Српски језик у (кон)тексту, зборник радова,* књ. 1, Крагујевац: ФИЛУМ. 407–427.

Ристић – Ивановић 2011: Ристић С., Ивановић Н., „Предлог за модернизацију рада на речнику САНУ", *Граматика и лексика у словенским језицима,* Нови Сад : Матица српска; Београд : Институт за српски језик САНУ. 529–553.

Saint-Dizier, – Viegas 2005: Saint-Dizier, P. (ed.) Viegas, E. (ed.), *Computational lexical semantics,* Cambridge: Cambridge University Press.

Schryver 2003: Schryver, G.-M. de, "Lexicographers' Dreams in the Electronic–Dictionary Age", International Journal of Lexicography, vol. XVI/2, Oxford: Oxford University Press. 143–199.

Фекете 1993: Фекете Е., *О Речнику српскохрватског књижевног и народног језика САНУ*, Сто година лексикографског рада у САНУ, Београд: САНУ. 21–49.

Stana Ristic, Tanja Samardzic,
Milena Jakic, Aleksandra Markovic,
Nenad Ivanovic
Institute for the Serbial Language SANU, Belgrade

SIGNIFICANCE OF DIGITALIZATION LANGUAGE RESOURCES OF SERBO-CROATIAN DICTIONARY OF LITERARY AND NATIONAL LANGUAGE FOR DEVELOPMENT OF SCIENCE AND PRESERVATION OF CULTURAL HERITAGE

Summary

Existing material for creating modern Serbo-Croatian Dictionary of literary and national language of the Serbian Academy of Arts and Sciences is a collection of about six million leaflets that contain information allowing to determine the meaning, usage patterns and grammatical features of words in Serbian language.

The materials were carefully collected over a period of almost a hundred years and today provides a rich and indispensable source of knowledge about the Serbian language, culture and even history. The possibility for storing and maintenance of the SAAS Dictionary in digital form is created with the development of computer technology.

Our work has several objectives: a) we would like to emphasize that the digitalization of linguistic resources of the SAAS Dictionary would help to the development of science and conservation of cultural heritage; b) identifying of specific procedures that the digitalization project should involve, c) defining the quality standards for digitalization. Our review includes an analysis of the currently used material vocabulary, examination of the possibilities for its computer processing, as well as an estimate of required investment.

We are trying to show that this project, which would allow easier access to resources, as well as automatic search, requires relatively small investment in regard to expected benefit, not only for the community of researchers that is dealing with standardization of Serbian language, but also for society in general.

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

930.85:004.9(082)
001.891:004(082)
007:004(082)

ДИГИТАЛНИ извори у друштвено-хуманистичким истраживањима / уредници Александра Вранеш, Љиљана Марковић, Гвен Александер. - Београд : Филолошки факултет Универзитета ; [Вичита] : Универзитет Емпориа ; Београд : Народна библиотека Србије, 2012 (Београд : Белпак). - 268 стр. : илустр. ; 24 cm. -
(Дигитализација културне и научне баштине, универзитетски репозиторијуми и учење на даљину : тематски зборник у 4 књиге ; #књ. #3 = Digitalision of Cultural and Scientific Heritage, University Repositories and Distance Learning : thematic edition of collected works in 4 volumes ; #vol. #3)

На спор. насл. стр.: Digital Sources in Social Studies and Humanities. - Тираж 500.
- Радови на срп. и енгл. језику. - Резимеи на срп. и енгл. језику. - Напомене и библиографске референце уз текст. - Библиографија уз већину радова.

ISBN 978-86-6153-108-8 (Књ. 3)
ISBN 978-86-6153-105-7 (за издавачку целину)
1. Вранеш, Александра [уредник]
a) Културна добра - Дигитализација - Зборници b) Научно-истраживачки рад - Зборници c) Информациони системи - Зборници
COBISS.SR-ID 193527308