Digitalizacija pisnega narečnega gradiva v dialektološki sekciji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU v Ljubljani

Peter Weiss in Andrejka Žejn

IZVLEČEK: V prispevku je predstavljeno optično branje (skeniranje) narečnega gradiva, ki od leta 2003 poteka v dialektološki sekciji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU v Ljubljani, in sicer obseg in oblika optično prebranega gradiva, dokumentiranje slik ter standardi in postopki pri optičnem branju.

ABSTRACT: The article describes the project of scanning the dialectal corpus material that has been going on since 2003 in the Dialectal section of the Fran Ramovš Institute of the Slovene Language SRC SASA in Ljubljana. Presented are the quantity and the form of the scanned material, the documenting of pictures, as well as the scanning standards and procedures.

1 V dialektološki sekciji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU v Ljubljani od februarja 2003 teče projekt, katerega izvedba se je pokazala kot nujno potrebna pri izdelovanju Slovenskega lingvističnega atlasa (SLA), ki se pripravlja, pa tudi pri Slovanskem lingvističnem atlasu (OLA), katerega izdelovanje (v sodelovanju z drugimi nacionalnimi komisijami) že poteka. Karmen Kenda-Jež (1999: 9) v pregledu obdobij dela za SLA ugotavlja, da se je z izdajo Vodnika po zbirki narečnega gradiva za Slovenski lingvistični atlas (SLA) Francke Benedik (Benedik 1999) končalo četrto obdobje nastajanja SLA in začelo peto, v katerem je ena od nalog tudi zasnova ustreznega načina računalniškega hranjenja in obdelave gradiva, kar se začenja ravno z omenjenim projektom.

Odločitev za skeniranje (tj. optično branje, računalniško zajetje slik, vendar brez optičnega prepoznavanja znakov) gradiva je bila sprejeta zato, ker je gradivo v računalniški obliki pri graditvi računalniške podatkovne zbirke lažje dostopno (tudi več uporabnikom hkrati) in obvladljivo, ker je podatke iz poskeniranega gradiva v računalniški zbirki lažje dokumentirati (na ustreznih mestih je npr. sorazmerno preprosto priložiti povezavo na sliko ustrezne izvirne strani), zato, da se izvirniki ne bi poškodovali, pa tudi zato, ker je kazalo, da se bo sekcija morala preseliti v druge, manjše prostore, nekateri sodelavci pa bi bili prisiljeni delati doma, pri čemer gradiva v klasični, papirni obliki ne bi bilo priporočljivo prenašati naokoli. Poleg tega je gradivo, ki ga hrani dialektološka sekcija, marsikdaj v enem samem izvodu,

večkrat pa se izvirnik oz. drugi, fotokopirani izvod (odgovori na vprašalnico za SLA) hrani le še v knjižnici Oddelka za slovenistiko in Oddelka za slavistiko Osrednje humanistične knjižnice (na Filozofski fakulteti) v Ljubljani. Dva izvoda, pri čemer je kopija praviloma slabše kakovosti ali pa kdaj vezana tako neustrezno, da je delo z njo prav težavno, če nista kopiji kar oba, sta vsekakor napredek v primerjavi z gradivom, ki je hranjeno v enem samem izvodu, vendar je možnost poškodovanja, izgube ali uničenja katerega od njiju še vedno velika. Edini izvod pa je zelo ranljiv, in če izgine, ga je le stežka, če ne kar nemogoče obnoviti. (Tako je Senahid Halilović s Filozofske fakultete v Sarajevu na zasedanju mednarodne komisije OLA oktobra 2004 v Bratislavi poročal, da se je med vojno v Bosni in Hercegovini izgubilo gradivo iz vseh 13 zapisovalnih točk za OLA iz Bosne in Hercegovine; ohranila se je le kopija odgovorov za eno točko, ki so jo imeli v Beogradu. Rokopisno gradivo so hranili v stavbi Akademije znanosti in umetnosti Bosne in Hercegovine v Sarajevu in vojna v zadnjem desetletju 20. stoletja je z očitno izgubo gradiva vknjižila in izterjala tudi ta davek. Tovrstne zapise je čez čas nemogoče obnoviti, sploh v krajih, ki jih je česalo in nacionalno »čistilo« nasilje, kot se je zgodilo v Bosni in Hercegovini. Tudi tipkopis pete knjige Slovarja slovenskega knjižnega jezika (SSKJ) je bil junija 1991, ko je Ljubljana doživela grožnjo z letalskim napadom, podobno ranljiv, čeprav je bil izdelan v kopiji. Zelo občutljiva je še vedno listkovna kartoteka v leksikološki sekciji inštituta za slovenski jezik ZRC SAZU v Ljubljani, ki vsebuje sijajno in neprecenljivo gradivo na več kot 6 milijonih enkratnih listkov, ki so bili uporabljeni pri izdelavi Slovarja slovenskega knjižnega jezika. Kot kažejo izkušnje s knjižničnimi katalogi, so novejši povsem elektronski (kot naš Cobiss), po katerih se da iskati na različne načine in jih sproti dopolnjujejo, starejši, listkovni katalogi pa so preprosto poskenirani in po njih se da ne ravno udobno iskati z listanjem po posameznih snopih tudi na internetu (npr. po katalogih Nacionalne knjižnice Češke republike v Pragi, prim. http://katif.nkp.cz/). Podobno pot bi kazalo ubrati pri inštitutski listkovni kartoteki, ki bi lahko bila kot predhodnica dobro internetno dopolnilo elektronskega inštitutskega korpusa, kakršen je Nova beseda – tako bi se zelo povečala uporabnost kartoteke, ki je marsikdaj zaradi čisto tehničnih ovir nedostopna, recimo ponoči, za osnovni namen, tj. SSKJ, pa je bila že izrabljena.)

2 Opis gradiva

2.1 Gradivo za SLA z opisi govorov in/ali odgovori na vprašalnico za SLA je rokopisno v obliki zvezkov formata A4 in A5 in listkov formata A6 ter tipkopisno formata A4 (Benedik 1999: 19, 118–127). Gradivo v zvezkih je, predvsem zaradi različnih zapisovalcev iz različnih obdobij, v različnih formatih. T. i. zvezkoteko, arhiv vseh zvezkov z zapisi sodelavcev sekcije in fotokopij izvirnikov študentskih diplomskih nalog, je uredila Francka Benedik v letih 1985–1986 (Kenda-Jež 1999: 8). Zvezki so izvirni zapisi ali fotokopije diplomskih nalog. Zapisi Tineta Logarja so običajno v rokopisnih zvezkih formata A4. V enakem formatu so še zapisi nekaterih drugih zapisovalcev, sodelavcev inštituta in zunanjih sodelavcev. Nekaj zapisov iz prve polovice osemdesetih let 20. stoletja je v posebnih zvezkih formata

A5, tj. v vprašalnici, ki je bila v sedemdesetih letih tehnično preurejena in sfotokopirana in ima predviden prostor za zapisovanje odgovorov (Benedik 1999: 17). Nekateri zapisi so nastali tudi v okviru magistrskih ali doktorskih del.

V dialektološki sekciji so dostopni tudi koncepti diplomskih nalog, ki jih je Tine Logar kot mentor za SAZU dobil od študentov za preverjanje gradiva za OLA in nasploh za delo na akademiji. Iz teh konceptov so seveda nastale naloge študentov, so pa vseeno dragoceno gradivo zaradi mentorjevih popravkov in komentarjev ob zapisih. V načrtu je tudi optično branje konceptov, ki bodo shranjeni v posebni mapi znotraj mape ustrezne točke.

Pri vezavi fotokopiranih diplomskih nalog ni bil puščen dovolj širok rob, da bi bila pri optičnem branju vedno vidna stran v celoti, poleg tega pa imajo nekatere naloge tudi popravke in komentarje Tineta Logarja, ki se na črno-beli fotokopiji ne ločijo dobro od zapisov študentov, zato so bili za optično branje zapisov, ki so nastali kot diplomske naloge, uporabljeni izvirniki diplomskih nalog s filozofske fakultete.

Tine Logar je dopisoval, običajno z rdečo barvo, komentarje in popravke tudi v svoje zapise, čemur so bile prilagojene nastavitve pri optičnem branju.

Diplomske in seminarske naloge so zapisane rokopisno ali tipkopisno, v zadnjem primeru (ter v magistrskih in doktorskih nalogah) večinoma z na roko dopisanimi posebnimi znaki zaradi pomanjkanja ustreznih znakov pri pisanju na pisalni stroj in v zadnjem času z računalnikom. Nekatere naloge so zapisane po fotokopirani prej omenjeni vprašalnici s prostorom, predvidenim za vpisovanje odgovorov. Format je A4, nekatere so v zvezku, druge na posameznih listih ali polah, novejše pa so vezane. Malokatera naloga vsebuje samo zapis gradiva po vprašalnici za SLA, večkrat so dodani še opis glasovnega sestava in oblikoslovja ter primer narečnega besedila.

Format listkov v t. i. listkoteki je A6. Za zdaj je torej optično prebran zapis na listkih za eno točko, ki je samo na listkih in še ni vložen v listkoteko. Pri zapisih na listkih gradivo ni razporejeno tako kot v zvezkoteki po mreži krajev, ampak »po zaporednih številkah vprašanj« in tako pri »tovrstnih zapisih ne moremo dobiti pregleda celotnega govora, če ne preiščemo celotne kartoteke« (Benedik 1999: 19). Pri nadaljnjem delu bo treba najti še ustrezen in učinkovit način za optično branje, urejanje in shranjevanje izredno obsežne listkovne kartoteke.

Zapis gradiva na posameznih listih v formatih A4, A5 in A6 omogoča optično branje s podajalnikom (v nasprotju z optičnim branjem gradiva v zvezkih ali na polah, kjer je treba položiti na skener vsako posamezno stran posebej ali po dve skupaj), pri čemer je treba posamezne datoteke preimenovati, saj jih program oštevilči oz. poimenuje zaporedno, kot so optično brane. Pri zvezkih formata A5 je treba datoteke še preurediti, in sicer je treba vse skene obrniti za 90 stopinj, kar omogoča lažje branje.

Delo se je začelo s skeniranjem gradiva, ki je samo v zvezkih. Po vodniku je vseh zapisov 473, od tega sta 102 samo v zvezkih in 85 samo na listkih, tako da je 286 zapisov po vprašalnici tako na kartotečnih listkih kot tudi v zvezkih. Skupno je po vodniku 388 zapisov v zvezkih. Doslej (do novembra 2004) so bili optično prebrani in urejeni vsi zapisi vprašalnice, ki so samo v zvezkih (101, brez golega

opisa), en zapis na listkih, pet golih opisov govorov, en zapis besedila, dva zapisa točk zunaj mreže in štirje zapisi po vprašalnici za OLA; ti so bili dodatek diplomskih nalog, ki so nastale kot zapisi po vprašalnici za SLA.

Prvih 14 razdelkov vprašalnice za SLA zajema 665 zaporedno oštevilčenih vprašanj, sledi razdelek Razno, ki so ga nekateri puščali praznega, pri drugih pa najdemo veliko podatkov o posebnostih narečja, ki so ga zapisovali. Številčenje v razdelku 16 (Gramatična vprašanja) se nadaljuje z zaporedno številko 700.

Zapisovanje gradiva poteka že od leta 1946, medtem pa se je deloma spreminjala tudi vprašalnica oz. številčenje. Prva je bila t. i. Ramovševa vprašalnica, ki je obsegala 835 oštevilčenih vprašanj, razdeljenih v 16 razdelkov. Prvih 14 razdelkov, t. i. leksikalni del vprašalnice, vsebuje vprašanja 1–665, sledi razdelek Razno, nato pa razdelek Gramatična vprašanja, oštevilčen s 700–780.

Ko so gradivo po vprašalnici za SLA začeli zapisovati študenti, je skušal Jakob Rigler vprašalnico, ki je bila namenjena enemu samemu jezikovno visoko razgledanemu raziskovalcu, prilagoditi tako, da bi lahko kljub večjemu številu raziskovalcev in njihovi manjši usposobljenosti dosegli primerljive izsledke (Kenda-Jež 1999: 7). Vprašanja, ki so zahtevala odgovor za več designatov, je poskušal največkrat razdeliti na A, B, C (npr. 51A, 51B, brez 51), kadar pa obstaja v knjižnem jeziku nadpomenka ali sopomenka, je vprašanje razčlenil na podvprašanja a, b, c (npr. 737, 737a, 737b).

Dopolnitve so iz časa, ko je bila zbrana že več kot polovica gradiva (Benedik 1999: 16–17). Po Riglerjevi vprašalnici je vprašanj skupaj sicer 880, vendar je zaradi načina številčenja dodatnih vprašanj oz. podvprašanj ohranjeno zaporedno številčenje od 1 do 870, kar kljub dopolnjeni vprašalnici omogoča enotno poimenovanje optično prebranih strani.

V nekem ne natančneje določljivem obdobju so nastali zapisi še po neki tretji vprašalnici, ki ni dokumentirana in ima drugačno številčenje (posamezna vprašanja glede na Ramovševo oz. Riglerjevo vprašalnico so združena pod eno zaporedno številko brez notranje razdelitve na podvprašanja). Pri zapisih po tej vprašalnici (doslej sta bila optično prebrana dva taka zapisa) je bila za označitev razdelka Razno uporabljena številka 644 namesto 666, tako da je razdelek Razno tudi tu uvrščen pred razdelek Gramatična vprašanja, ki se namesto s 700 začne s številko 647. Odgovori so oštevilčeni z zaporednimi številkami od 1 do 817, brez kakršne koli razdelitve znotraj vprašanj.

V zapisih za SLA torej najdemo tri različna številčenja, ki so upoštevana v imenih datotek (skenov):

- po prvotni, Ramovševi vprašalnici,
- po preurejeni, Riglerjevi vprašalnici,
- po vprašalnici, ki se konča pred razdelkom Razno s 646 (verjetno še neka tretja vprašalnica, mogoče Kolaričeva).

2.2 Gradivo za OLA je rokopisno (odgovori na vprašalnico – Voprosnik OLA 1965) v zvezkih formata A5 in A4 in tipkopisno formata A4 (večina zbirnih seznamov, pripravljenih iz odgovorov za slovensko jezikovno področje), vse to pa je lahko opremljeno s pripisi z barvnim (kemičnim) svinčnikom, kar je treba upoštevati pri določanju ločljivosti in nastavitev pri skeniranju. Doslej so bili

poskenirani odgovori iz dveh točk za OLA in več registratorjev zbirnih seznamov. Kot je bilo že rečeno, obstajajo štirje zapisi po vprašalnici za OLA, od tega trije za točke zunaj mreže OLA.

2.3 Nekaj je poskeniranega tudi iz knjig (npr. Voprosnik OLA 1965; OLA – Vstupitel'nyj vypusk 1994; FO 1981), in sicer zato, ker so nekatere knjige težko dosegljive (npr. vprašalnica za OLA), druge pa so težke za prenašanje (npr. t. i. sarajevski Fonološki opisi). Ta dela bodo na voljo na spletni strani OLA (http://OLA.zrc-sazu.si). Poskenirani so tudi nekateri slovenski narečni slovarji (npr. Šašel – Ramovš 1936–1938 in rokopisni Rožanski besednjak Josipa Šašla iz leta 1957), pa tudi posamezne magistrske naloge in disertacije članov dialektološke sekcije. Za lažje delo so bili ti skeni kot izvedenke pretvorjeni v format PDF (pripona datotek .pdf), ki ga je mogoče brati s programom adobe acrobat reader. Ker ravno člani mednarodne in slovenske nacionalne komisije OLA veliko potujejo v različna evropska mesta, neprimerno lažje jemljejo s seboj na pot nekaj več gigabajtov podatkov na trdem disku prenosnega računalnika kot knjige, čeprav so za kabinetno delo knjige seveda še vedno potrebne.

3 Strojna oprema

Za samo skeniranje je bil uporabljen osebni računalnik z operacijskim sistemom windows 2000 z 256-megabajtnim pomnilnikom in 50-gigabajtnim trdim diskom ter skener HP, ki omogoča zajemanje z največjo ločljivostjo 1200 pik na palec, s podajalnikom listov, ki se je med delom izkazal za zelo koristen pripomoček in smiselno naložbo, saj pri ustreznem gradivu prihrani veliko časa. Jeseni 2003 smo pridobili še 200-gigabajtni zunanji trdi disk, na katerem je shranjeno gradivo, ki ga občasno prekopiramo tudi na cedeje oz. devedeje, tako da se datoteke ne bodo izgubile, tudi če bi prišlo do izgube katerega od nosilcev.

4 Nastavitve

Rokopisno gradivo se skenira v ločljivosti 300 pik na palec v 256 sivinskih odtenkih, le gradivo, ki vsebuje popravke in dopolnila v večinoma rdeči barvi, je skenirano v 256 (= 2^8) barvah, iz česar potem vedno lahko dobimo prostorsko varčnejše sivinske odtenke (medtem ko poti v drugo smer ni). Izkazalo se je, da je tolikšna ločljivost zadostna, hkrati pa ob zdajšnjem stanju računalniške strojne in programske opreme omogoča dovolj hitro obdelavo slik in tudi paketno obdelavo. – Tipkopisno in natisnjeno gradivo je poskenirano v ločljivosti prav tako 300 pik na palec v dveh barvah (beli in črni).

Slike strani so kot osnova shranjene v formatu TIFF (datoteke imajo pripono .tif) s stiskanjem LZW, ki ne povzroči izgube podatkov. Stiskanje LZW je primerljivo s standardom za stiskanje ZIP, ki prav tako ne povzroči izgube podatkov. Na ta način je dobljeni rezultat optimalen, saj so slike tako dobre, da jih lahko na računalniškem zaslonu v dvomnih primerih uporabimo povečane, za kar bi pri papirju

morali uporabiti lupo. Format JPEG (pripona datotek je .jpg) omogoča veliko stiskanje, pri njem datoteke zasedajo manj prostora (uporablja se za izmenjavo datotek po internetu), povzroči pa izgubo podatkov, kar seveda ne zagotavlja povrnitve v izvirno stanje. V določenih primerih je vendarle smiselna paketna pretvorba večje količine datotek v ta format, recimo ko želimo zmanjšati obseg datotek in smo ob tem pripravljeni žrtvovati kakovost slik.

Primer: datoteka »0000-03 -- Uvod« (kar pomeni eno od datotek z uvodom) iz točke 247 (Ribnica) za SLA je bila poskenirana v ločljivosti 300 pik na palec v formatu TIFF in shranjena v stisnjeni obliki LZW, in ker gre za rokopis, v 256 sivinskih odtenkih (od katerih jih je v tej datoteki uporabljenih 237). Stran je velika 21,6 × 29,7 cm, tj. 2550 × 3510 slikovnih pik. Nestisnjena izvirna datoteka bi obsegala 8,54 MB, stisnjena v obliki LZW zaseda 3,58 MB in se shrani. Iz nje je potem mogoče narediti izvedenke, pri katerih se izgubi del podatkov, vendar pa so tudi manj obsežne, v različnih formatih, kot je JPEG. Pri shranjevanju v tem formatu lahko med drugim izbiramo med najboljšo in najslabšo kakovostjo (tj. med 100 in 1), tako da je obseg datoteke v kakovosti 100 3,51 MB, v kakovosti 50 376 KB in v kakovosti 1 70 KB. Pri stiskanju LZW v formatu TIFF datoteka še vedno zaseda 42 % prvotne velikosti (kot rečeno: brez izgube podatkov), v formatu JPEG pa pri izbiri najslabše kakovosti le še 0,8 % – podatki na sliki so vidni, vendar lahko koga motijo kvadratne sive zaplate. V formatu JPEG je smiselno izbrati kaj od tistega, kar je blizu najboljši kakovosti in ne preveč potratno s prostorom pomnilnika – če je to sploh še kaka težava.

Pri paketni pretvorbi (npr. iz formata TIFF v format JPEG ali pri obračanju strani za 90 ali 180 stopinj) se je izkazal odlični program irfanview avtorja Irfana Škiljana (http://www.irfanview.com), za paketno preimenovanje (številčenje ipd.), ki je uporabno pri razvrščanju datotek s posameznimi poskeniranimi stranmi v knjigi, pa je primeren prav tako zastonjski program lupas rename 2000 avtorja Ivana Antona Albarracina (http://rename.lupasfreeware.org).

5 Poimenovanje datotek

Slike iz vsakega zvezka imajo v računalniku svojo mapo, poimenovano s SLA ali OLA, in številko kraja, ki jo ima zapisana točka v vsakokratni mreži. Tako je na primer mapa za kraj Kobarid, ki ga v mreži točk za SLA najdemo pod številko 70, poimenovana »SLA070 Kobarid«. Kjer je za posamezno točko več zapisovalcev, so posamezni zapisi poimenovani po njih. Tako so v mapi »SLA404 Gornji Senik -- Felsőszölnök« za zdaj (ko so optično prebrani šele zapisi, ki so samo v zvezkih, ne pa še zapisi na listkih in v zvezkih) mape, poimenovane z letnico in imenom zapisovalca, tj. »1976 Marija Kozar-Mukič«, »1983 Marija Bajzek« in »1986 Marija Magdolna Horvat«. Do datoteke z želenim poskeniranim odgovorom, npr. na vprašanje 184 *sosed* po vprašalnici za SLA, se da priti zelo hitro, v največ štirih korakih: SLA > SLA404 Gornji Senik -- Felsőszölnök > 1983 Marija Bajzek > 0174--0194.tif.

Zaradi narave gradiva je bilo treba vzpostaviti poseben sistem poimenovanja

datotek. Upoštevati je bilo treba, da gre za začetek večjega projekta, v katerega ne bo vključeno le skeniranje zvezkovnega gradiva za SLA, poleg tega pa je vsakič optično prebran celotni zapis, od naslovnice do zadnje strani. Tako za SLA kot za OLA imajo v imenih računalniških datotek številke zapisovalnih točk tri številke (v SLA je točk 406, v OLA 853), številke odgovorov pa po štiri (v SLA jih je sicer samo 870, v OLA pa jih je 3454). Tako npr. v imenu datoteke SLA001 pomeni zapisovalno točko 1 v SLA, SLA0001 pa odgovor na vprašanje 1 iz vprašalnice za SLA. (Čez čas bo za kake druge projekte treba poiskati drugačne, prijaznejše načine poimenovanja zapisovalnih točk s slovenskega jezikovnega ozemlja, t. i. govoreča imena, v katerih bo prva črka pomenila narečno skupino, druga narečje v tej narečni skupini, tretja in četrta črka bosta simbol kraja, temu bo pri krajih zunaj Slovenije dodana še kratica države, za vezajem pa bodo lahko sledile začetnice informatorja: ŠzSK-FI torej pomeni štajerska narečna skupina, zgornjesavinjsko narečje, Spodnje Kraše (temu v SLA ustreza številka 314), informator Franc Irmančnik (1908–1991), KzBr(A) pa pomeni koroška narečna skupina, ziljsko narečje, kraj Brdo – Egg, država Avstrija (v SLA je to točka 1).)

Datoteke z optično prebranimi stranmi z zapisom same vprašalnice so označene po številkah odgovorov na posamezni strani. Ime datoteke s sliko konkretne strani zajema številki mejnih odgovorov, ki sta na strani: prva številka je številka odgovora, ki se na konkretni strani začne ali nadaljuje, druga številka je zadnji odgovor na strani, ne glede na to, ali se odgovor konča na isti strani ali pa se nadaljuje še na naslednji. Iz tehničnih razlogov med obema štirištevčnima poljema namesto pomišljaja stojita stična vezaja.

Začetne strani z naslovnico imajo oznako 0000, ker pa jih je običajno več, so označene kot 0000-01, 0000-02 itd. Po podobnem načelu so poimenovane končne strani, tj. strani, ki sledijo zapisu oštevilčene vprašalnice, z 0999-01, 0999-02 itd. Posebno enoto, kot je bilo že omenjeno, predstavlja razdelek 15, imenovan Razno, ki »izstopa iz celotnega koncepta, saj sploh ne vsebuje vprašanj, ampak samo dve navodili za zapisovanje. To je načrtovani (še neizdelani) del.« (Benedik 1999: 16) Ravno zaradi te (še) neizdelanosti je ta del, ki je sicer velikokrat izpuščen, njegov obseg pa je lahko od ene besede do več strani, označen z 0666 (zadnje vprašanje pred razdelkom Razno je oštevilčeno s 665, prvo v naslednjem razdelku pa s 700) oz. 0666-01, 0666-02 itd., razen že navedene izjeme.

Nekatera vprašanja zahtevajo navajanje različnih oblik, zato so odgovori zapisani na več straneh; v takih primerih je uporabljeno podobno načelo razvrščanja oz. poimenovanja kot pri uvodnih in končnih straneh ter razdelku Razno, in sicer z zaporednim oštevilčenjem za številko odgovora (npr. pri vprašanju 577 *visok*, ki zahteva celotno sklanjatev v določni in nedoločni obliki v vseh treh spolih ter primernik in presežnik v vseh treh spolih, so običajno pri zapisih, ki navajajo vse zahtevane oblike, strani označene z 0577-01, 0577-02 itd.). Pri takem številčenju bo uporabnik gradiva vedno že iz poimenovanja posameznih datotek lahko ugotovil, ali se odgovor nadaljuje na naslednji strani.

Ponekod se zapisovalci niso dosledno držali številčenja iz vprašalnice; tako so na primer črke ob zaporednih številkah izpuščene ali pa pride do zamenjave malih črk, ki označujejo podvprašanja in so v vprašalnici zapisane brez presledka

za številko in zaklepaja (130a *podstrešna soba*), in tistih, ki označujejo razdelke posameznega vprašanja in pri katerih je za črko še zaklepaj (125 *peta*; a) *na nogi*, b) *na čevlju*). Ponekod sta zamenjani velika in mala črka. Pri shranjevanju oz. poimenovanju datotek je bilo, če se je le dalo, upoštevano številčenje po vprašalnici. Pri tem pridejo v poštev seveda samo mejni odgovori na straneh.

Zaradi večje preglednosti, kaj posamezna datoteka poleg vprašalnice vsebuje in kje oz. zaradi hitrejšega dostopa do iskanih podatkov imajo spremljevalne strani izpolnjene vprašalnice, označene kot 0000 (začetne strani), 0999 (končne strani) in 0666 (Razno), za nestičnima vezajema nakazano vsebino. Pomembnejše navedene vsebine so tele:

- naslovnica (»0000 -- Naslovnica« oz. »0000-XX -- Naslovnica« – XX tule nadomešča konkretno zaporedno številko, ki je vedno zapisana z dvema številkama, saj je v drugih primerih kdaj spremljevalnih strani nekaj deset, nikoli pa jih ni več kot 99); stran, ki vsebinsko običajno ni informativna in je označena ravno zato, da uporabnik ve, da je pri iskanju določenih podatkov ni treba odpirati;
- komentar Tineta Logarja pri diplomskih nalogah o (ne)zanesljivosti gradiva 0000-XX -- Komentar TL; v enem primeru doslej je datoteki dodana še optično prebrana stran s komentarjem Francke Benedik na naslovnici fotokopije diplomske naloge (»0000-XX -- Komentar FB«) o (ne)zanesljivosti gradiva, ki je zapisan samo na fotokopiranem zvezku, ne pa tudi na izvirniku;
- informatorji (»0000-XX -- Informatorji« ali »0999-XX -- Informatorji«);
- zapis besedila (»0000-XX -- Zapis besedila« ali »0999-XX -- Zapis besedila«);
- pri opisu govora, ki je lahko na uvodnih ali končnih straneh: glasoslovje (kjer je mogoče, posebej samoglasniki in soglasniki), oblikoslovje, naglas, zaimki, vezniki, prislovi, členki ... (npr. »0000-XX -- Zaimki« ali »0999-XX -- Soglasniki«);
- v razdelku Razno so dodane oznake pri tistih zapisih, kjer je gradivo vsebinsko enotno, npr. hišna imena, ledinska imena, orodje, domača obrt idr. (npr. »0666-XX -- Hišna imena«).

Dodatno komentiranje v imenih datotek – ta v novejših različicah operacijskega sistema windows lahko obsegajo do 255 znakov – je zaradi dolžine na videz nepraktično, vendar pa uporabniku gradiva omogoča boljši pregled nad gradivom in hitrejši dostop do iskanih vsebin.

V posebnem protokolu, ki bo urejen kot podatkovna zbirka, bo navedeno, po kateri vprašalnici je narejen zapis, tako da bo lahko uporabnik gradiva ob še drugih podatkih iz protokola prej našel določeni odgovor oz. glede na vprašalnico ugotovil, kje je kak odgovor. Poleg tega bo že iz protokola razvidno ime optično prebrane datoteke, koliko zvezkov je za posamezni govor, kdo so zapisovalci, letnica zapisa, oblika in format (zvezek, listi, pole, diplomska naloga; rokopis, tipkopis), vsebina (vprašalnica, opis govora, dodatno gradivo, npr. zapis hišnih ali ledinskih imen, zapis besedila), kdo in kdaj je skeniral ter kje se hrani papirno in elektronsko gradivo.

6 In naprej? Najprej bo treba dokončno poskenirati in urediti zvezke z zapisi

za SLA in OLA, potem listkovno kartoteko za SLA, nakar pridejo na vrsto monografije (tudi objav sodelavcev sekcije, in sicer za morebitno predstavitev na internetu) in starejše dialektološke objave. Digitalizirati bo treba tudi drugo gradivo, ki se hrani v dialektološki sekciji, in sicer predvsem zvočne posnetke, ki se zdaj hranijo na magnetofonskih trakovih, in redke starejše fotografije, ki so bile posnete pri terenskem zapisovanju predvsem za SLA. Vse to je delo, ki ga bo težko opraviti v doglednem času.

Navedenke

Benedik 1999 = Francka Benedik, *Vodnik po zbirki narečnega gradiva za Slovenski lingvistični atlas (SLA)*, Ljubljana, Založba ZRC, 1999.

FO 1981 = *Fonološki opisi srpskohrvatskih/hrvatskosrpskih, slovenačkih i makedonskih govora obuhvaćenih Opšteslovenskim lingvističkim atlasom*, Sarajevo, ANUBiH, 1981 (Posebna izdanja LV, Odjeljenje društvenih nauka 9).

Kenda-Jež 1999 = Karmen Kenda-Jež, Predgovor, v: Francka Benedik, *Vodnik po zbirki narečnega gradiva za Slovenski lingvistični atlas (SLA)*, Ljubljana, Založba ZRC, 1999, str. 5–9.

Šašel – Ramovš 1936–1937 = Josip Šašel – Fran Ramovš, Slovar, v: Josip Šašel (zbral) – Fran Ramovš (priredil), *Narodno blago iz Roža*, Maribor, Zgodovinsko društvo, 1936–1937 (Arhiv za zgodovino in narodopisje II), str. 101–122.

Voprosnik OLA 1965 = *Voprosnik Obščeslavjanskogo lingvističeskogo atlasa*, Moskva, Nauka, 1965.

OLA – Vstupitel'nyj vypusk 1994 = *Obščeslavjanskij lingvističeskij atlas – vstupitel'nyj vypusk: Obščie principy, spravočnye materialy*, Moskva, Nauka, [2]1994.

The Digitalization of the Written Dialectal Corpus Material in the Dialectal Section of the Fran Ramovš Institute of the Slovene Language SRC SASA in Ljubljana

Summary

Since 2003 the scanning of the handwritten, typewritten and printed dialectal corpus material has been going on in the Dialectal section of the Fran Ramovš Institute of the Slovene Language SRC SASA in Ljubljana. The corpus material was first intended for the ongoing and future projects (Slavic linguistic atlas, Slovene linguistic atlas) and was kept in notebooks, on individual sheets of paper and in the form of card files. The first step was the scanning of the notebook archives, i.e. with the materials that were kept only in notebooks. This project is necessary in order to preserve the originals and to protect valuable information from loss or destruction during work processes. In the era of computers it is even easier to work with scanned pictures and other materials, e.g. for the documenting in data bases. Another benefit

is that the digitalized materials are accessible to several users at the same time. For the whole scanning procedure standards had to be set in order to guarantee quality pictures and straightforward documenting, together with the use of logical file names; at the same time the limitations in file size and the saving of data had to be taken into account.