Simon Krek
Institut "Jožef Stefan"

SSKJ v slovarski bazi

V okviru slovenskega leksikografskega prostora so na področju opisa splošnega jezika v preteklih letih potekali vzporedni leksikografski projekti, katerih rezultat sta Leksikalna baza za slovenščino (projekt Sporazumevanje v slovenskem jeziku) kot osnova za Predlog za izdelavo Slovarja sodobnega slovenskega jezika (SSSJ) ter Slovar novejšega besedja slovenskega jezika (SNB) kot slovar v tiskani obliki, katerega vsebina dopolnjuje obstoječi Slovar slovenskega knjižnega jezika in bo predvidoma integriran v drugo izdajo SSKJ. Rezultat prvega projekta predstavlja možno osnovo za izdelavo digitalne slovarske baze oz. slovarja, katerega primarna oblika je digitalna ("*born digital dictionary*"), rezultat drugega pa je knjižno zasnovan slovar, kakršnega poznamo iz evropske leksikografske tradicije 19. in 20. stoletja. V prispevku bomo predstavili nekatere elemente v mikro- in makrostrukturi slovarjev, ki nastajajo primarno v digitalni obliki in se posvetili razmisleku o možni integraciji podatkov iz Slovarja slovenskega knjižnega jezika v večfunkcionalno digitalno slovarsko bazo.

Dictionary of Standard Slovenian in a dictionary database

In Slovenia, there were two parallel lexicographic projects in recent years, one resulting in the Slovene Lexical Database (project Communication in Slovene) as the basis for the Proposal for the compilation of the Dictionary of modern Slovene and the other in the Dictionary of newer Slovene vocabulary as a printed dictionary whose contents represents an addition to the existing Dictionary of Standard Slovene and will be integrated in its second edition. The result of the first project represents a possible basis for the compilation of a dictionary or a dictionary database whose primary form is digital (born digital dictionary) while the result of the second is a dictionary made for print that is known from the European lexicographic tradition from the 19th and 20th century. The article presents some elements in micro- and macrostructure of the born digital dictionaries and reflects on the possible integration of the data from the Dictionary of Standard Slovene into a multi-purpose digital dictionary database.

Ključne besede: slovarska baza, tiskani slovar, retrodigitalizirani slovar, slovar z digitalno zasnovo

Keywords: dictionary database, printed dictionary, retrodigitized dictionary, born digital dictionary

1 Uvod

Pri razmisleku o tem, kako zastaviti nov slovenski slovarski projekt, se sredi drugega desetletja 21. stoletja kot pomembna strateška odločitev postavlja vprašanje, ali je novi slovar v smislu notranje strukturiranosti še vedno treba zastaviti na podlagi tiskane zasnove SSKJ, ali je bolje zasnovati slovarsko bazo na novo na tak način, da je izvorni tiskani medij v celoti opuščen, kar pomeni, da je struktura baze bolj podobna računalniškim bazam podatkov kot pa linearno potekajočemu besedilu v tiskanem mediju. Ta razmislek ni samo tehnične narave, odločitev je bistveno povezana tako z vsebino slovarja, organizacijo slovarskega dela, predvsem pa s prihodnjo predvideno rabo slovarske baze v najrazličnejše namene. Pomembno je poudariti, da slovenščina ni edini jezik, ki se je na področju leksikografije pri prehodu v digitalno dobo znašel pred omenjeno dilemo, zato si bomo ogledali tudi nekatere rešitve, ki izhajajo iz tujih praks, predvsem Veliki slovar poljskega jezika, Slovar sodobnega nizozemskega jezika ter Danski slovar.

2 Zgodovina na kratko

Uporaba računalnikov v leksikografiji ima že dolgo zgodovino, ki se začenja nekje v 60. letih prejšnjega stoletja. V poročilu o prvem večjem računalniškem slovarskem podvigu urednik slovarja *The Random House Dictionary of the English Language* pravi: "Dobro se spomnim svojega prvega srečanja z računalniki: leta 1959 sem pri založbi Random House delal na novem slovarju in prišlo mi je na misel, da bi bil računalnik idealen za razvrščanje in obdelavo različnih vrst podatkov, za izdelavo katerih sem bil odgovoren" (Urdang 1984: 152).[1] Rezultati računalniške vaje so bili uspešni: "Kodiranje informacij na različnih ravneh – geslo, iztočnica, izgovor, razlaga/e, variantne oblike, etimologija, podiztočnice, ilustracije – in več kot 150 področij, katerim so bile lahko pripisane definicije – botanika, kemija, računalništvo itd., so omogočili pripravo podatkov na vseh nivojih in neodvisno za vsako področje, kar je zagotavljalo bolj poenoteno obravnavo in veliko večjo konsistentnost med

[1] Za referenco se zahvaljujem dr. Dušanu Gabrovšku z Oddelka za anglistiko in amerikanistiko Filozofske fakultete Univerze v Ljubljani.

povezanimi deli podatkov, kot je bilo to mogoče doseči pri drugih slovarjih" (ibid: 155–156). Končni rezultat je bila baza podatkov: "Na kratko, ustvarili smo zmogljivo bazo podatkov, izdelano iz slovarja s približno 260.000 gesli, ki smo jo lahko preiskovali po mili volji" (ibid: 156).[2] To je bilo leta 1966, dve leti po tem, ko je bil izdan poskusni snopič Slovarja slovenskega knjižnega jezika in štiri leta pred izidom prve knjige A–H.

Drugi slovarski projekt, ki ga je smiselno omeniti v kontekstu vprašanja organiziranosti slovarskih podatkov, je *Longman Dictionary of Contemporary English (LDOCE)* iz leta 1978. Baza omenjenega slovarja je bila namreč kmalu po izdaji tiskanega slovarja uporabljena za številne sekundarne namene, predvsem za potrebe računalniškega procesiranja angleščine. V obsežni študiji (Boguraev in Briscoe 1984) so tako opisani načini, kako je bila leksikalna baza (*lexical database*) LDOCE, kot so jo tudi imenovali, uporabljena za sestavljanje leksikonov besednih oblik in različnih besednih seznamov, gradnjo taksonomij in ontologij, strojno skladenjsko razčlenjevanje, semantično procesiranje, raziskave kolokativnosti, računalniško sintezo govora in druge naloge.[3] To je omogočila stukturiranost baze podatkov v leksikalni bazi, ki je bila organizirana po t. i. vozliščih (*nodes*) oz. samostojnih strukturiranih delih s specifičnimi podatki: o iztočnici in njenih oblikah, izgovorjavi, naglaševanju, zlogovanju, skladenjskih vzorcih, semantičnih kategorijah, področnih omejitvah, definicijah itd. (glej Sliko 1).

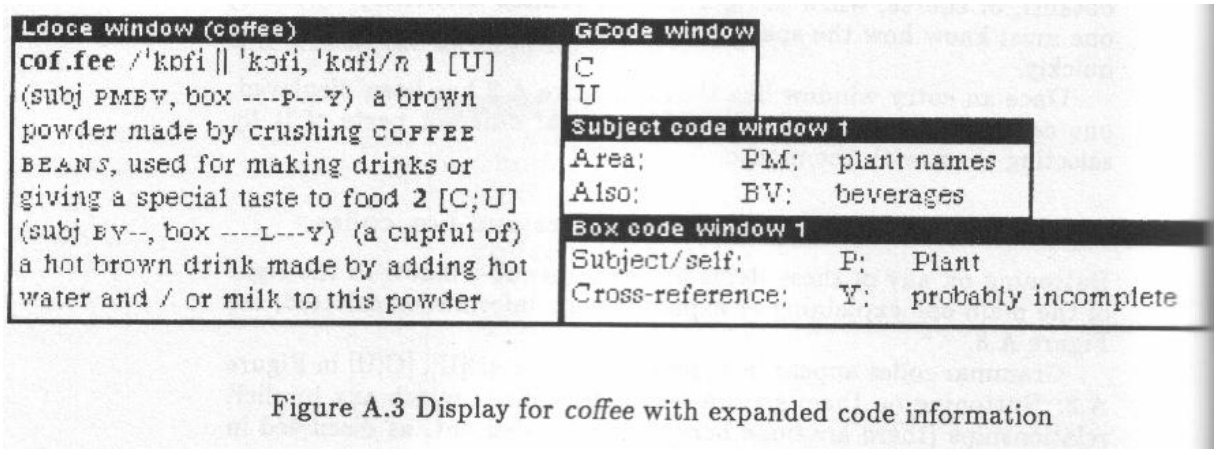


Figure A.3 Display for *coffee* with expanded code information

Slika 1: Leksikalna baza LDOCE

Za angleški jezik bi torej lahko rekli, da sedemdeseta, še bolj pa osemdeseta leta prejšnjega stoletja že zaznamuje prehod slovarjev v digitalno okolje, tako pri uporabi korpusnih virov za

[2] Omenjeni slovar oz. slovarska baza je bila kasneje uporabljena tudi za izdelavo črkovalnika *The Random House ProofReader*, ki je bil izdan leta 1982.

[3] Nikakor ni naključje, da sta pri tej publikaciji sodelovala dva pomembna avtorja, ki se ukvarjata s področjem umetne inteligence: Bran Boguraev (IBM, projekt Watson) in Yorick Wilks (University of Sheffield).

slovarje (projekt COBUILD),[4] urejanju slovarjev in izkoriščanju za namene procesiranja naravnih jezikov (*Natural Language Processing*).

V tem času v Sloveniji še sestavljamo SSKJ, ki nastaja v klasični obliki: četrta knjiga Preo–Š izide 1985, zadnja knjiga z iztočnicami T–Ž leta 1991. Slovar v tistem času – razen zadnje, pete knjige – ne obstaja v digitalni obliki in ga ni mogoče izkoriščati v drugačne namene, bodisi za računalniško procesiranje ali za kompleksnejše jezikoslovne analize. Ena od sodelujočih pri procesu digitalizacije leta tako 1993 zapiše:

> "Glede na velik interes vseh, ki se ukvarjajo s pisanjem v slovenščini in z različnimi raziskavami slovenskega jezika ter pri svojem delu uporabljajo računalnike, je izdaja Slovarja v računalniški obliki nujna. /.../ Ko nam je uspelo pritegniti k delu računalniškega strokovnjaka mag. Primoža Jakopina, smo se odločili za optični prenos z bralnikom slike. Ministrstvo za znanost in tehnologijo nam je leta 1992 odobrilo namenska sredstva za nakup računalnika (PC 486) in bralnika slike, Ministrstvo za kulturo pa sredstva za nakup programske opreme za razpoznavanje besedila (Lecturus). Jakopin je svoj računalniški program EVE tako izpopolnil, da se je jeseni 1992 začelo poskusno delo." (Hajnšek Holz 1993: 421)

Digitalizacija SSKJ poteka od 1992 do 1994 za potrebe stavljenja in izdaje SSKJ v eni knjigi (izid leta 1994) in za kasnejše izdaje v elektronski obliki na disketah (1997).

3 Retrodigitalizirani slovarji in slovarske baze

V prej omenjenih angleških zgledih se zrcali dilema, s katero so se spopadali tako rekoč vsi slovarji v devetdesetih letih prejšnjega stoletja. Oba omenjena angleška slovarja sta že nastajala v digitalni obliki, kar je omogočalo veliko boljši in konsistentnejši nadzor in izkoriščanje vsebine, četudi je bila splošni javnosti ta potem dostopna v tiskani obliki. Precej drugačne pa so dileme pri t. i. retrodigitaliziranih slovarjih, ki izhodiščno niso nastajali v digitalni obliki, temveč so bili naknadno digitalizirani – v angleškem prostoru je najbolj tipičen primer *Oxford English Dictionary*.

Za razliko od angleščine, ki je pri rabi računalnikov v leksikografiji prehitevala za desetletje ali dve, je v devetdesetih letih prejšnjega stoletja potekala množična digitalizacija slovarjev iz tiskane osnove tudi pri drugih jezikih. Kot kaže zgled LDOCE, pri slovarjih gre za močno

[4] Slovar COBUILD je bil izdan leta 1987, temeljil pa je na analizi 7,3-milijonskega besedilnega korpusa z istim imenom.

strukturirano besedilo, ki ga lahko uporabimo tudi za najrazličnejše druge namene, če je baza konsistentno notranje organizirana. Iz tega izhaja, da je ključna odločitev pri procesu digitalizacije, ali bo nastali retrodigitalizirani slovar upošteval predvsem elemente oblikovanja, torej različnih vrst ali slogov pisav, ali pa bodo v tem procesu upoštevane tudi vsebinske informacije in notranja strukturiranost makrostrukture slovarja (identifikacija geselskih člankov), predvsem pa njegovih mikrostrukturnih delov (npr. identifikacija razlag, oznak, zgledov ...). Ker pri tiskanih slovarjih, ki so bili izdelani brez računalniških urejevalnikov, ni bilo niti možno niti smiselno v celoti slediti strogi notranji strukturiranosti mikrostrukturnih podatkov, je še bolj kritično vprašanje, ali bodo v procesu digitalizacije nekonsistentnosti odpravljene oz. ali se bomo tega vprašanja sploh lotili, ali pa bo uporabljena struktura dovolj ohlapna, da bo dovoljevala nekonsistentnosti, s tem pa bomo potencialno ogrozili ali omejili uporabo baze za druge namene.

V času izdelave LDOCE (1978) še ni bilo univerzalnih standardov, s pomočjo katerih bi lahko oblikovanje povezali s strukturo vsebine. Ta potreba je bila na svetovni ravni močno zaznana v osemdesetih letih, ko je najprej nastal standard SGML (*Standard Generalized Markup Language*), ki mu je kasneje sledil XML (*eXtended Markup Language*). Standard SGML je bil objavljen leta 1986 in je bil torej v času digitalizacije SSKJ že na voljo, žal pa ga Inštitut za slovenski jezik pri procesu digitalizacije ni uporabil za strukturiranje slovarske baze. Sistem, ki je bil uporabljen, je v glavnem upošteval oblikovanje, poleg samih geselskih člankov pa so bili označeni tudi nekateri mikrostrukturni elementi, ki jih je bilo mogoče konsistentno prepoznavati, denimo iztočnice, oznake, zgledi in podobno. Kodiranje SSKJ v takratni obliki je prikazano na Sliki 2 (začetek gesla "maček").[5]

```
<F1>m<030>ček<A10>1
<F2><->čka<SP><A4>
<F16><#76>
<F2><SP><A1><(><030><) >
<F3>1<. >
<F6>mačji<SP>samec<: >
<F7>maček<SP>leži<SP>na<SP>peči<; >naš<SP>maček<SP>rad<SP>lovi<SP>miši<;
>maček<SP>mijavka<, >praska<, >prede<; >črn<SP>maček<;
>urno<SP>kakor<SP>maček<SP>je<SP>splezal<SP>na<SP>drevo<; ><A4>
```

Slika 2: Struktura v Slovarju slovenskega knjižnega jezika

[5] Slika izhaja iz verzije baze SSKJ, ki je bila leta 1998 objavljena na prosto dostopnem delu spletne strani Inštituta za slovenski jezik.

Naslednja možnost pretvorbe SSKJ v format SGML/XML je bila na voljo v letih 1998–2000, ko je za slovarje v našem geografskem oz. jezikovnem prostoru potekal projekt CONCEDE (*Consortium for Central European Dictionary Encoding 1998–2000*), v okviru katerega je bil izdelan standard SGML/XML za strukturiranje bolgarskih, čeških, estonskih, madžarskih in romunskih enojezičnih slovarjev, ne pa tudi slovenskega. Institutu »Jožef Stefan«, ki je sodeloval pri tem projektu, Inštitut za slovenski jezik Frana Ramovša namreč ni dal na voljo že obstoječe digitalizirane baze SSKJ niti ni želel sodelovati v projektu.

Pretvorba SSKJ v format XML je tako potekala šele v letih 2011–2013, torej 20 let po digitalizaciji, in sicer za potrebe uvoza v slovarski urejevalnik iLex in s tem povezane druge izdaje SSKJ oz. izdelave Novega slovarja slovenskega jezika (Ledinek, Perdih 2012a, 2012b). Navedki iz prispevkov sodelavcev Inštituta za slovenski jezik, ki so se ukvarjali s pretvorbo v format XML, kažejo, da se na Inštitutu pravzaprav zavedajo slovarske realnosti v 21. stoletju:

> »Morda najodločilneje je sodobno leksikografijo zaznamovalo dejstvo, da leksikografi in uporabniki slovarskih priročnikov ne dojemajo več kot (izhodiščno) knjižnih jezikovnih virov, ampak kot večnamenske razširljive strukturirane računalniško berljive podatkovne baze, v katerih so podatki ustrezno hierarhizirani, (standardno) označeni in medsebojno povezani. V letu 2011 je bila oblikovana tudi XML-shema za nastajajoči Novi slovar slovenskega jezika, enojezični razlagalni slovar v obsegu približno 70.000 gesel, ki naj bi bil nekoliko manj ambiciozen naslednik Slovarja slovenskega knjižnega jezika.
>
> Pri njeni pripravi smo se srečali z izzivom, kako vzpostaviti XML-shemo, ki bo omogočala ohranjanje leksikografske tradicije predhodnika v segmentih, ki so se izkazali za dobre in ki so jih uporabniki vajeni, in sicer tudi na ravni izmenljivosti podatkov med obema bazama, hkrati pa omogočala vzpostavitev novih leksikografskih praks, ki so se kot ustrezne potrdile v praksi sodobne, tudi tujejezične leksikografije, pri čemer naj bi bila shema oblikovana čim bolj striktno in preudarno, tj. tako, da v čim večji meri preprečuje nesistematično interpretiranje podatkov, hkrati pa njihovo predstavitev na uporabniku čim bolj prijazen način v elektronski obliki.« (Ledinek, Perdih 2012a: 128)

Težav pri pretvorbi SSKJ v format XML iz obdobja 2011–2013 je več. Prva je ta, da niti sama shema XML, torej formalni zapis slovarske strukture, niti vzorčna gesla v formatu XML niso bila javno objavljena. Tako pravzaprav ne moremo oceniti, do katere mere držijo trditve, da

bo slovarska baza »omogočala vzpostavitev novih leksikografskih praks«. Avtorjema lahko zgolj verjamemo na besedo. Druga težava je, da avtorja omenjata zgolj povezljivost »podatkov med obema bazama«. V času, ko se evropski leksikografi sprašujejo o množičnem povezovanju leksikografskih baz v tako rekoč neskončno mrežo podatkov o različnih jezikih,[6] je omenjeni cilj bistveno premalo ambiciozen. Prva stvar, o kateri bi bilo treba razmišljati že na ravni slovenščine, ne glede na univerzalno povezljivost z drugimi jeziki, je konsistentno povezovanje obstoječih in bodočih digitalnih virov za slovenski jezik: korpusov, leksikonov besednih oblik, pravopisnih priročnikov, dvo- ali večjezičnih slovarjev, digitalnih knjižnic itd.

Še najbolj pa je pri načrtovanem slovarju oz. slovarjih problematično, da do sedaj edini objavljeni primeri gesel iz Novega slovarja slovenskega jezika (Snoj 2012: 96–101) kažejo, da je zasnova še vedno izrazito knjižna. Slovar bo imel zelo zapleten sistem oznak, kazalk in drugih slovarskih sredstev, ki so znani iz sveta tiskanih slovarjev, njegovo zasnovo bodo v veliki meri določali izračuni velikosti fontov, dolžine vrstic in podobno. V nadaljevanju prikazujemo gesla iz slovarja, kakor so predstavljena v omenjenem prispevku:

máčka -e $ž_{S200g}$ **1.** domača žival, ki lovi miši: *Mačka prede* **2.** mačja samica: *Mačka ima mladiče* **3.** kot povdk., nav. s pril., ekspr. izraža, da se v osebku imenovani osebi ženskega spola pripisuje, da je postavna, privlačna: *Ona je dobra mačka* || s pomenskim dopolnilom kot delovalnik *Pri mizi sta sedeli dve črnolasi mački* **máčkin** -a -o svoj. prid.$_{49c}$ k pomenoma 1.–2. ↑**máčica** -e $ž_{S200g}$ manjš. **máčkica** -e $ž_{S200g}$ manjš. **T** žival., mn. družina živali roparic z okroglo glavo, gibčnim trupom in ostrimi kremplji; *Felidae*: *levi, tigri, gepardi in druge mačke* **E** = hrv., srb. *máčka*, slovaš. *mačka* < slovan. **mačьka* iz vabnega klica *mac(a)*

máčica -e $ž_{S200g}$ **1.** manjš. od mačka 1.–2.: *mačice in kužki* **2.** nav. mn. socvetje v obliki podolgovate kepice: *vrbove mačice* || vejice s takimi socvetji: *šopek mačic* **3.** nav. s pril., ekspr. ljubka, mikavna ženska: *spoznati ljubko mačico* **E** < **maččica*, manjš. od ↑máčka

pomaránča -e $ž_{S200g}$ **1.** južni sadež z oranžno lupino: *limone in pomaranče* || zveza *grenka pomaranča* sorta pomaranč, primerna za sokove, marmelade **2.** pog. pomarančevec 1. **pomaránčica** -e $ž_{S200g}$ manjš. **E** ◁ star. it. *pomarancia* iz *pomo* 'jabolko, sadež' + *arancia* 'pomaranča' ◁ špan. *naranja* ◁ arab. *nārandž* ◁ perz. *nārandž*

[6] Primera sta denimo evropski projekt COST *European Network of e-Lexicography* (http://www.cost.eu/domains_actions/isch/Actions/IS1305) z enim od ciljev, da se razišče »panevropska narava večine besedišča evropskih jezikov«, ter delavnice na konference Euralex 2014 *Publishing and consuming lexicographical resources in the linked (open) data cloud* – Objava in uporaba leksikografskih virov v povezanem oblaku (odprtih) podatkov.

nosoróg -a $_{m S1k}$ velika čokata žival z zelo debelo kožo in enim ali dvema izrastkoma na sprednjem delu glave; *Rhinocerus*: *povodni konji in nosorogi* **nosorógov** -a -o svoj. prid.$_{P1k}$ **nosorógec** -gca $_{m S1k}$ manjš. **nosoróginja** -e $_{ž S200f}$ **nosoróginjin** -a -o svoj. prid. $_{P1k}$ **E** ↑nós + ↑róg po zgledu lat. *rhīnoceros* iz gr. *rhís* 'nos' + *kéras* 'rog'

dekàn -ána in dekán -a m člov. $_{S1a}$ in $_{S1e}$ **1.** predstojnik fakultete: *Dekan je sklical sejo akademskega zbora* **2.** predstojnik dekanije: *Škof je napovedal pogovore z dekani* **dekánov** -a -o svoj. prid. $_{P1a}$ **dekánica** -e $_{ž S200a}$ k pomenu 1. **dekáničin** -a -o svoj. prid. $_{P1a}$ **dekánja** -e $_{ž S200a}$ k pomenu 1. **dekánjin** -a -o svoj. prid. $_{P1a}$ **dekánka** -e $_{ž S200a}$ k pomenu 1. **dekánkin** -a -o svoj. prid. $_{P1a}$ **E** ◁ nem. *Dekan* ◁ lat. *decānus* 'desetnik' k *decem* 'deset'

lisják -a $_{m S1e}$ **1.** lisičji samec: *V daljavi je zalajal lisjak* **2.** kot povdk., ekspr. izraža, da se v osebku imenovani moški osebi pripisuje, da je zvita, prebrisana: *Naš sosed je lisjak* || kot psovka *ti lisjak stari* || s pomenskim dopolnilom kot delovalnik *Temu lisjaku ne zaupaj!* **lisjákov** -a -o svoj. prid. $_{P1a}$ **E** < **lisъjakъ* k ↑lisíca

Slika 3: Gesla v Novem slovarju slovenskega jezika

Del o vprašanju (retro)digitalizacije in s tem povezanimi načrtovanimi dejavnostmi lahko sklenemo z ugotovitvijo, da je digitalizacija SSKJ potekala sočasno z digitalizacijo slovarjev za druge evropske jezike, da pa je njena nadaljnja usoda vključevala dve izraziti težavi: (a) baza ni bila ustrezno strukturirana, saj se ni nahajala v bazi podatkov s konsistentno notranjo strukturo (primer LDOCE) oz. v formatu SGML/XML, in (b) baza ni bila dana na voljo raziskovalni in širši skupnosti za namene raziskovanja, računalniškega procesiranja itd. Glede na sicer skope podatke Inštituta za slovenski jezik pa lahko sklepamo, da bodo tudi načrtovani slovarji zasnovani za tiskani medij.

4 Slovarji z digitalno zasnovo in Leksikalna baza za slovenščino

Slovarja *The Random House Dictionary of the English Language* in *Longman Dictionary of Contemporary English* smo izpostavili kot primera slovarjev, ki sta bila za razliko od retrodigitaliziranih slovarjev že zelo zgodaj zasnovana kot slovarski bazi, vendarle pa sta bila tudi tadva primarno izdelana za tiskani medij. V zadnjem desetletju pa so svetovni trendi, kot je širjenje spleta in njegova vsesplošna, tudi mobilna dostopnost, splošni prehod v digitalne medije, tudi z radikalnim opuščanjem tiska, ter hiter razvoj informacijsko-komunikacijskih tehnologij močno posegli tudi na področje tradicionalne leksikografije. Začel se je prehod s

slovarjev, izdelanih za tisk, na *born digital dictionaries* – slovarje, ki so v osnovi zasnovani za digitalni medij in jih ne utesnjujejo tradicionalne omejitve tiskanega medija v smislu prostora ali razporeditve podatkov na (tiskani) strani ter izkoriščajo povezljivost tako na spletu kot tudi med različnimi bazami (npr. korpusi) in znotraj slovarske baze. Na strokovni ravni se indikativni trendi zadnjih let kažejo v nastanku in popularnosti serije konferenc *E-lexicography in the 21st century* (2009, 2011, 2013) ter vzpostavitvi vseevropskega projekta COST *European Network of e-Lexicography* (2013–2017), pri katerem denimo sodelujejo predstavniki 25 držav, večinoma z inštitutov za nacionalne jezike.

V nadaljevanju bomo izpostavili tri jezike oz. slovarje, ki so značilni za omenjeni trend: danščina z Danskim slovarjem (*Den Danske Ordbog*), poljščina z Velikim slovarjem poljskega jezika (*Wielki słownik języka polskiego*) ter nizozemščina s slovarjem sodobnega nizozemskega jezika (*Algemeen Nederlands Woordenboek*).[7]

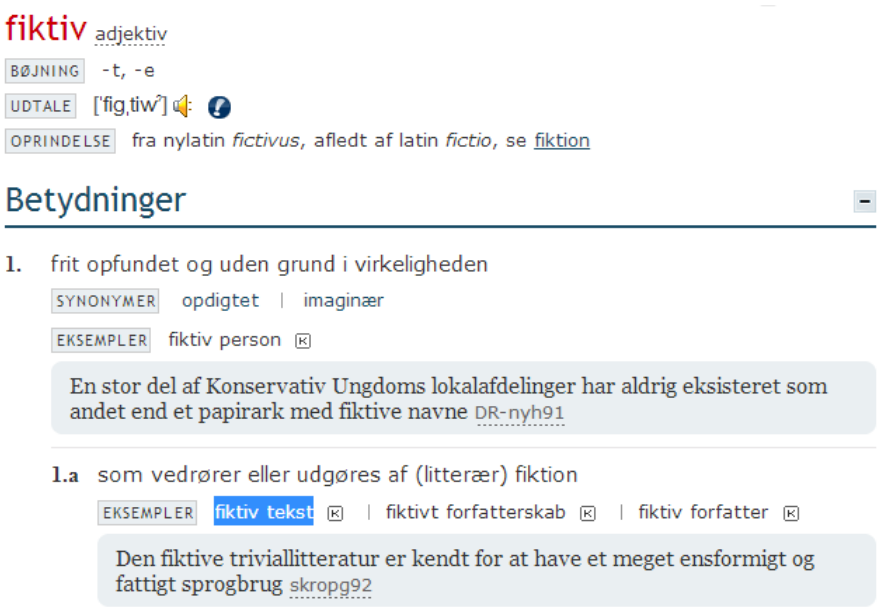
Danski slovar izpostavljamo kot predstavnika linije, ki skuša kombinirati izvorno tiskani slovar z drugimi bazami, pri čemer so različni podatki bodisi vključeni ali povezani s slovarsko bazo. V nadaljevanju prikazujemo del gesla *fiktiv* (sl. fiktiven, navidezen) iz Danskega slovarja v formatu XML:

```
<Artikel EntryID="11012881">
    <Iddel>
      <ID></ID>
      <Holem>fiktiv</Holem>
      <Lemklas>adj.</Lemklas>
    </Iddel>
    <BøjFon>
      <Bøjdel>
        <Bøjning>
          <Bform>
            <Norm>
              <txt>-t</txt>
            </Norm>
          </Bform>
          <Bform>
            <Norm>
```

[7] Spletne strani: http://ordnet.dk/ddo/, http://www.wsjp.pl/, http://anw.inl.nl/.

```
            <txt>-e</txt>
          </Norm>
        </Bform>
      </Bøjning>
    </Bøjdel>
    <SpFon/>
    <Fondel>
      <Komfon>
        <Fon Lydfil="11012881_1">HfigBtiwZ</Fon>
      </Komfon>
    </Fondel>
  </BøjFon>
    <...>
</Artikel>
```

Struktura gesla kaže sestavo, ki je značilna za slovarje, izhajajoče iz tiskanega medija, s številnimi hierarhičnimi nivoji za razmeroma preproste tipe informacij. Poleg tega se tiskana osnova kaže tudi z ohranjanjem okrajšanih delov gesla. Isto geslo je na spletu vizualizirano takole:



Slika 4: Danski slovar

Kot vidimo, je osnovna besednovrstna opredelitev iztočnice v bazi denimo navedena z okrajšavo, vizualizirana pa je z izpisom, skupaj s povezavo na slovnično pojasnilo. Baza vsebuje tudi podatke, ki spletnim uporabnikom (še) niso dostopni. Tak primer je povezava posameznega pomena v slovarju z bazo danskega WordNeta oz. DanNeta:

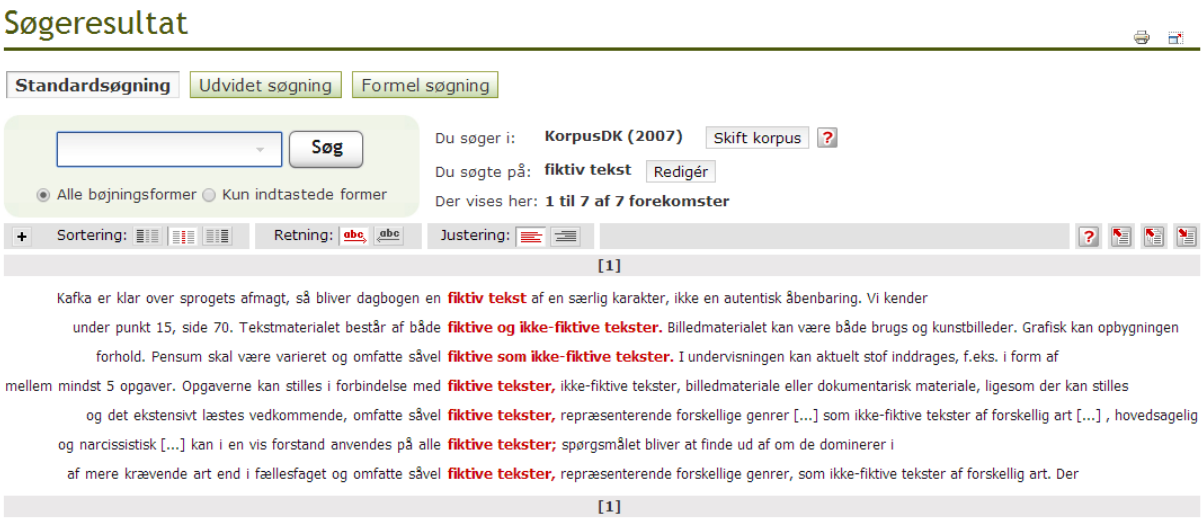
```
<Semdel SemID="70020247">
  <Semem>
      <Denbet DanNetSemID="21019448" DanNetSemType="Semem">frit
      opfundet og uden grund i
      virkeligheden<Genprox>opfundet</Genprox></Denbet>
    <...>
  </Semem>
</Semdel>
```

Poleg dodanih nevizualiziranih podatkov slovarska baza vsebuje tudi povezave na zunanje baze, na primer na danski korpus. Kot primer navajamo zvezo *fiktiv tekst*, ki je v strukturi XML zabeležena na naslednji način:

```
        <Kerne>
    <txt>fiktiv tekst</txt>
    <RelDelt>
      <Position>
        <Form>fiktiv</Form>
        <Exthenv>
          <ReferRel>
            <ArtRef>¤fiktiv, adj.</ArtRef>
            <EntryRef>11012881</EntryRef>
          </ReferRel>
        </Exthenv>
      </Position>
      <Position>
        <Form>tekst</Form>
        <Exthenv>
```

```
            <ReferRel>
               <ArtRef>¤tekst, sb.</ArtRef>
               <EntryRef>12000522</EntryRef>
            </ReferRel>
         </Exthenv>
      </Position>
   </RelDelt>
</Kerne>
```

Spletni uporabniki zgornji slovarski zapis v formatu XML vidijo kot povezavo na korpus (glej sliko zgoraj), ki ob kliku prikaže naslednji rezultat:

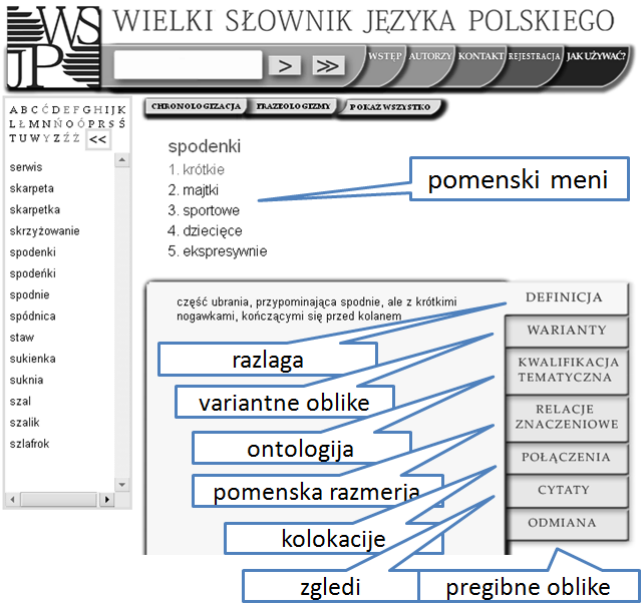


Slika 5: Danski korpus

Na opisani način je torej slovarska baza izvorno tiskanega slovarja nadgrajena in povezana z drugimi digitalnimi bazami.

Ostala dva slovarja – poljski in nizozemski – sta bila že na izhodišču zastavljena drugače, in sicer eksplicitno kot slovarja, ki sta primarno namenjena uporabi v digitalnem okolju, zato je tudi njuna zasnova in posledično shema XML povsem drugačna. Najbolj izstopajoča značilnost teh slovarjev je organiziranost podatkov v strogo hierarhično strukturo, kjer uporabnik do podatkov dostopa preko t. i. pomenskega menija, podatki na ostalih nivojih pa so organizirani po posameznih v sebi zaključenih enotah, ki so na podoben način tudi vizualizirani na spletu:

Slika 6: Slovar sodobnega nizozemskega jezika (geslo *koe*, sl. krava)



Slika 7: Veliki slovar poljskega jezika (geslo *spodenki*, sl. spodnjice, kratke hlače)

Podatki v omenjenih slovarjih so torej organizirani neobremenjeno z linearno organizacijo besedila v tiskanih slovarjih, vsebujejo pa lahko tudi neomejeno količino informacij in povezav na druge jezikovne vire.

Na podoben način kot poljski in nizozemski slovar je organizirana Leksikalna baza za slovenščino (LBS) (Gantar 2009), izdelana v okviru projekta Sporazumevanje v slovenskem jeziku.[8] Oblikovana je kot mreža medsebojno povezanih leksikalnogramatičnih podatkov, ki so organizirani v šest nivojev. Notranja hierarhična ureditev temelji na semantičnem izhodišču, kar pomeni, da so podatki na posameznem nivoju podrejeni pomenskim lastnostim

[8] Spletna stran: http://www.slovenscina.eu/.

besede. Hierarhično najvišja je lema, tj. iztočnica v osnovni obliki, ki zastopa vse pripadajoče leksikalne enote, kamor so všteti posamezni pomeni in podpomeni, stalne besedne zveze in frazeološke enote.

Na pomenskem nivoju so zabeleženi osnovni pomeni in podpomeni obravnavane besede v iztočnici, ki so opredeljeni s pomenskimi indikatorji. Indikatorji so primarno namenjeni oblikovanju pomenskega menija, ki služi uporabniku za hitro navigacijo po večpomenskem geslu. Drugi del pomenske informacije predstavlja pomenska shema, ki se teoretično približuje pomenskim shemam, kot jih predvideva projekt FrameNet.[9]

Čeprav so skladenjski podatki v obliki osnovnega stavčnega vzorca za vsak posamezni pomen vključeni že v pomensko shemo, je eksplicitno skladenjskim podatkom namenjen skladenjski nivo. Na tem nivoju so za vsak registrirani pomen besede v iztočnici zabeležene skladenjske strukture in skladenjske zveze, pri glagolih glede na njihovo vlogo stavčnega organizatorja pa tudi stavčni vzorci.

Naslednji nivo je kolokacijski: na primer, skladenjska struktura *gbz Inf-GBZ* (glagol + nedoločnik) je zapolnjena s kolokacijami kot *[uspeti, poskušati, skušati, znati] omrežiti*, vzorec *kdo omreži koga* in *omrežiti s čim* pa s kolokacijami *[ženska] omreži*, *omrežiti [moškega, srce]* in *omrežiti s [čari]*. Vloga korpusnih zgledov, ki so navedeni na samostojnem nivoju, je ponazoriti in potrditi vse predhodne informacije ter hkrati pokazati obnašanje leksikalne enote v njenem najbolj naravnem in tipičnem okolju.

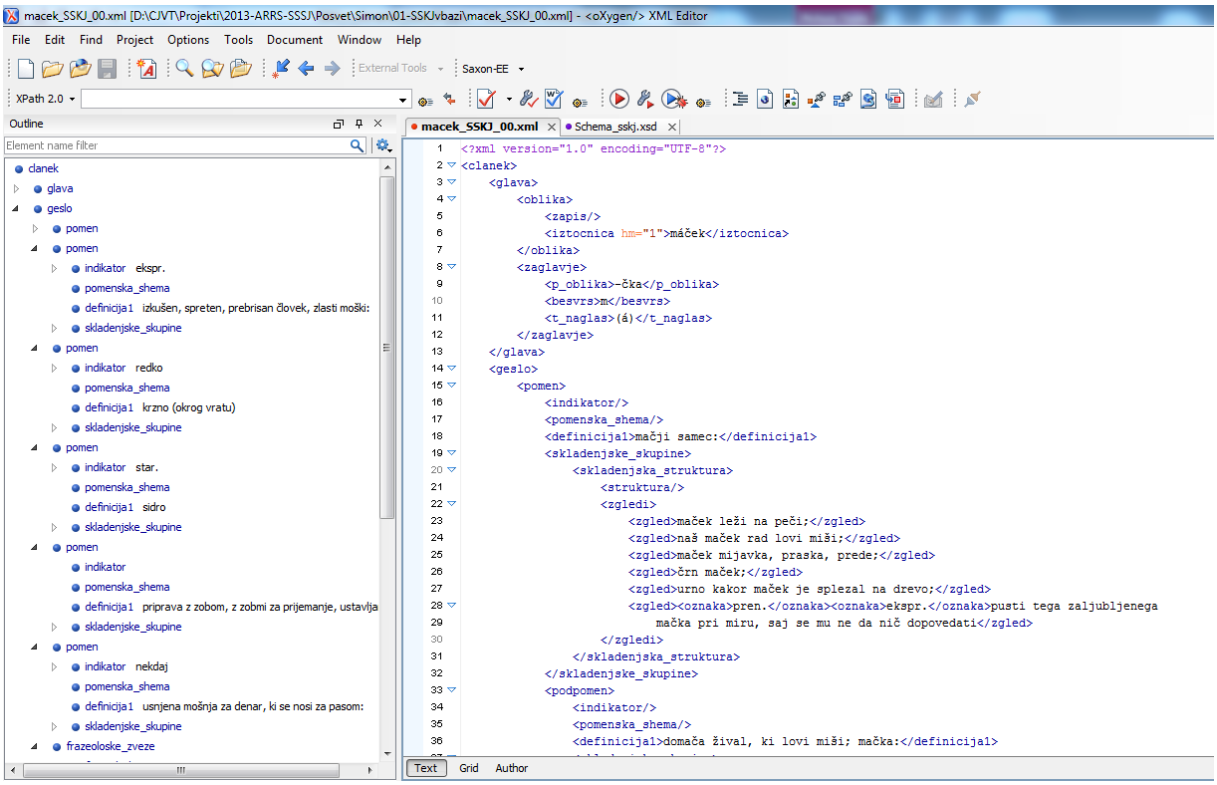
Stalne zveze in frazeološke enote so v LBS obravnavane kot samostojne leksikalne enote. Prve so vključene pod posamezni pomen ali podpomen besede v iztočnici, druge pa na koncu geselskega članka v samostojnem razdelku. Vsaka stalna zveza in frazeološka enota je opredeljena s pomenskim indikatorjem, lahko ima izkazane različne variantne oblike in tipično kolokabilno okolje ter mora biti potrjena z zgledi iz korpusa.

Leksikalna baza za slovenščino torej predstavlja osnovo za slovar slovenskega jezika z digitalno zasnovo, kakršen je bil konceptualiziran v Predlogu za izdelavo Slovarja sodobnega slovenskega jezika.[10]

Za potrebe prispevka je bil izveden tudi poskus vnosa podatkov srednje zahtevnega gesla iz SSKJ (»maček«) v strukturo LBS, z namenom, da preverimo možnost integracije podatkov,

[9] Spletna stran: https://framenet.icsi.berkeley.edu/fndrupal/.
[10] Spletna stran: http://www.sssj.si/.

kakor so organizirani v leksikalni bazi, in podatkov v obstoječem slovarju SSKJ. Poskus je bil izveden v orodju Oxygen, kakor prikazuje slika X:



Slika 8: Geslo "maček" v orodju Oxygen

Poskus je pokazal, da tehnično vključitev ni problematična, saj je bilo za integracijo vseh podatkov iz SSKJ v vsebinsko smiselne elemente LBS treba dodati zgolj nekaj elementov v zaglavju. Razlog za ta dodatek je v zasnovi LBS, ki predpostavlja, da se podatki o pregibnih oblikah, izgovoru itd. nahajajo v drugih bazah (npr. leksikonu besednih oblik), s katerimi je iztočnica povezana, niso pa del same baze. Poleg tega je bilo treba dodati tudi t. i. terminološko gnezdo, saj zasnova LBS predpostavlja, da so terminološke enote bodisi vključene v pomensko strukturo ali pa so samostojna gesla v bazi. Ostale dele gesla v SSKJ je bilo mogoče vključiti na logična mesta v LBS.

To pomeni, da je vsebinsko vključitev podatkov iz SSKJ v LBS načeloma možna, za končno oceno smiselnosti združitve podatkov pa bi bilo treba predvsem poznati obstoječo strukturo XML druge izdaje SSKJ oz. Novega slovarja slovenskega jezika.

5 Zaključek

Kakšna bo povsem formalna zasnova slovarja, ki ga bomo uporabljali v prihodnje, je odločitev, ki presega zgolj preprosto tehnično vprašanje izdelave sheme XML za obstoječi slovar. Odločitev vključuje niz dilem, katerih narava je v temelju leksikografska. Kolikor predpostavljamo, da bo prihodnji slovar eden od osrednjih virov v mnogo širšem ekosistemu jezikovnih virov za slovenščino, ki je povezljiv in povezan tako z leksikonom besednih oblik, leksikalno bazo, različnimi korpusi (govorjenega, pisnega jezika), bazami posnetega govora, slovarjem znakovnega jezika, slovenskim WordNetom (sloWNet), slovensko Wikipedijo in mnogimi drugimi, tudi dvo- ali večjezičnimi bazami, je smiselno zasnovati novi slovar kot hierarhično in fleksibilno organizirano slovarsko bazo, ne pa kot slovar s knjižno zasnovo.
Poskus integracije podatkov iz obstoječega SSKJ v strukturo Leksikalne baze za slovenščino kaže, da je te podatke mogoče vključiti in s tem omogočiti bodisi njihovo vizualizacijo na enotnem slovarskem portalu ali uporabo delov pri sestavljanju novega slovenskega slovarja.

Bibliografija

BOGURAEV, Bran, BRISCOE, Ted (ur.), 1989. *Computational Lexicography for Natural Language Processing*. London, New York: Longman.

GANTAR, Polona, 2009. "Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. " *Jezik in slovstvo* 54(3/4). 69–94.

HAJNŠEK-HOLZ, Milena, 1993. "Leksikografski problemi prenosa knjižne oblike Slovarja slovenskega knjižnega jezika v računalniško". *Jezik tako in drugače*. 420–432.

LEDINEK, Nina, PERDIH, Andrej, 2012a. "Izdelava XML-shem za slovarske projekte na primeru nastajajočih tipološko raznovrstnih slovarjev". *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012: zbornik 15. mednarodne multikonference Informacijska družba – IS 2012, zvezek C*. Institut »Jožef Stefan«, Ljubljana. Str. 123–128.

LEDINEK, Nina, PERDIH, Andrej, 2012b. "Uporaba XML-formata v leksikografiji na primeru oblikovanja XML-sheme za Slovar sinonimov slovenskega jezika". *Jezikoslovni zapiski: zbornik Inštituta za slovenski jezik Frana Ramovša*, 18/1. ZRC SAZU, Ljubljana. Str. 157–176.

SNOJ, Marko, 2012. "Podgesla v Novem slovarju slovenskega jezika". *Škrabčevi dnevi 7 – Zbornik prispevkov s simpozija 2011*. Nova Gorica.

URDANG, Laurence, 1984: A lexicographer's adventures in computing. *Dictionaries: Journal of the Dictionary Society of North America 6.1*. 150–165.