Uporaba XML-formata v leksikografiji na primeru oblikovanja XML-sheme za Slovar sinonimov slovenskega jezika

Nina Ledinek – Andrej Perdih

Cobiss: 1.02

Prispevek pojasnjuje, zakaj se je XML-format uveljavil kot standardni format za večnivojsko hierarhično strukturiranje jezikovnih podatkovnih zbirk in kako se s pomočjo XML-sheme nadzoruje formalna struktura in vsebina elementov v slovarski podatkovni zbirki. Prikazani so različni vidiki – tj. leksikografski oz. vsebinski vidik, praktični vidik ter tehnični vidik –, ki jih je pri strukturiranju kompleksnejših slovarskih podatkovnih zbirk v XML-formatu smiselno upoštevati. Sprejemanje odločitev je ponazorjeno s primerom oblikovanja XML-sheme za Slovar sinonimov slovenskega jezika.

Ključne besede: leksikografija, XML, XML-shema, slovar sinonimov slovenskega jezika

Using XML Format in Lexicography: Creating an XML Schema for the *Dictionary of Slovenian Synonyms*

This article explains why XML format has become established as the standard format for multilevel hierarchical structuring of linguistic databases and how an XML Schema can be used to manage the formal structure and content of elements in a dictionary database. Various aspects that must be taken into account when structuring complex dictionary databases in XML format are presented: the lexicographic or content aspect, the practical aspect, and the technical aspect. Decision-making is illustrated with the example of designing an XML Schema for the *Dictionary of Slovenian Synonyms*.

Keywords: lexicography, XML, XML Schema, dictionary of Slovenian synonyms

0 Uvod

Kljub temu da se jedrne naloge leksikografov pri oblikovanju slovarjev v samem bistvu od začetkov organiziranega slovaropisja pa vse do danes pravzaprav niso radikalno spremenile – ključne faze leksikografskega dela med drugim ostajajo zbiranje in priprava ustreznega gradiva, ki je izhodišče za analizo jezikovnih podatkov, oblikovanje geslovnika, analiza besedilnih zgledov in njihova pomenska ter slovnična, zlasti skladenjska, razčlemba, hierarhizacija interpretiranih podatkov na mikro- in makrostrukturni ravni itd. –, so se na področju leksikografije v zadnjih

petindvajsetih letih na tehnološki in konceptualni ravni zgodili ključni premiki, ki so odločilno vplivali na metodologijo sodobnega leksikografskega dela, uporabljeno tehnologijo in vzpostavljene standardne formate ter na dojemanje in posledično na uporabo slovarskih priročnikov.

Znanilec nekakšne leksikografske revolucije tako v tehnološkem kot konceptualnem smislu je bil sredi osemdesetih let prejšnjega stoletja leksikografski projekt COBUILD (prim. Sinclair 1987), v okviru katerega je bil oblikovan slovar Collins COBUILD English Language Dictionary, prvi slovar, ki je v celoti nastal na podlagi (elektronskega) referenčnega korpusa. Z vidika pričujočega prispevka je poleg dejstva, da je omenjeni projekt odločilno vplival na razvoj korpusov in orodij za obdelavo korpusnih podatkov, predvsem pa na zavest o pomembnosti uporabe korpusov v leksikografiji, najpomembneje, da je bil projekt COBUILD eden prvih leksikografskih projektov, pri katerem so slovarsko gradivo obdelovali s pomočjo specializiranih računalniških leksikografskih orodij, ključen pa je tudi podatek, da so redaktorji nastajajočo leksikalno bazo kot (pred)slovarsko besedilo dojemali kot večnamensko strukturirano računalniško berljivo bazo jezikovnih podatkov, zapisano v standardnem in preprosto berljivem ASCII-formatu (prim. Clear 1987: 51).

Prav tehnološki napredek na področju računalništva, katerega zametki so se kazali že v času cobuildovske revolucije, pa je dejavnik, ki je najodločilneje vplival na leksikografijo na prehodu v novo tisočletje, in sicer tako z vidika oblikovanja slovarskih priročnikov kot tudi njihove uporabe. Pomembno je na sodobno leksikografijo vplivalo zlasti dejstvo, da si leksikografi slovarja najpogosteje ne predstavljajo več kot (izhodiščno) knjižnega jezikovnega priročnika, ampak kot strukturirano razširljivo računalniško berljivo podatkovno zbirko, ki je uporabna za različne namene in v kateri so vsi podatki ustrezno hierarhizirani, (standardno) označeni in medsebojno povezani.

Objava podatkov na spletu in v obliki drugih elektronskih medijev omogoča bistveno drugačno izrabo slovarskih priročnikov, kot smo je bili vajeni nekoč. Tradicionalna mikrostruktura slovarjev se je zato, vsaj z vidika njihovih uporabnikov, nekoliko razrahljala, saj so uporabniku poleg »običajnega« slovarskega besedila vedno pogosteje na voljo dodatne (multimedijske) vsebine,[1] hkrati pa so se tudi meje med posameznimi slovarskimi (in drugimi jezikovnimi) podatkovnimi zbirkami nekoliko zabrisale. Pripravljavci slovarjev namreč svoje jezikovne vire vedno pogosteje objavljajo v obliki enovitih strukturiranih podatkovnih zbirk (npr. v okviru internetnih portalov), pri čemer lahko uporabniki svoje jezikovne zadrege rešujejo z brskanjem po različnih priročnikih in drugih virih hkrati. Omeniti velja še dejstvo, da elektronski medij v osnovi omogoča, da uporabnik prikaz jezikovnih podatkov vsaj deloma prilagodi svojim željam in potrebam, seveda če to omogoča sama strukturiranost podatkovne zbirke.

Spremembam na ravni danes običajne recepcije slovarskih priročnikov in drugih jezikovnih virov se je moral predhodno prilagoditi tudi sam redakcijski proces, v okviru katerega postaja vedno bolj običajno, da so podatki za konkreten slovarski priročnik zbrani v več medsebojno povezanih zbirkah podatkov, pri čemer ob

[1] Pogosto lahko uporabnik vzpostavljene rešitve tudi komentira, s čimer, gledano dolgoročno, postaja tudi njihov aktivni sooblikovalec.

objavi konkretnega priročnika v različnih oblikah želene podatke povežemo oz. jih avtomatsko izvozimo na ustrezna mesta. Vse omenjeno je seveda mogoče le, če so slovarske podatkovne zbirke in drugi z njimi povezani viri zapisani v obliki, ki zagotavlja jasno strukturiranje podatkovnih tipov, razširljivost izhodiščne podatkovne zbirke, preprosto shranjevanje podatkov ter njihovo prenosljivost med različnimi orodji. Kot standardni format za zapis jezikovnih podatkov se je zaradi svoje fleksibilnosti, univerzalnosti in zmogljivosti v zadnjem času uveljavil XML (eXtensible Markup Language).

1 Uporaba XML-formata v leksikografiji

Slovarji lahko vsebujejo veliko število pomensko različnih jezikovnih podatkov, zato imajo navadno razmeroma kompleksno mikrostrukturo. V knjižnih izdajah se to grafično kaže z različnim oblikovanjem besedila, odstavki in ločili med različnimi slovarskimi podatki. Že preprost slovar vsebuje številne podatke, kot so iztočnica, obrazila in izgovorjava, glede na vrsto slovarja pa lahko nastopajo še pomenski podatki, zgledi, podgesla, sinonimi, etimološki podatki, podatki o prvi pojavitvi, če naštejemo le nekatere izmed možnih. Za ustrezno računalniško obravnavno morajo biti vsi podatki v slovarju ustrezno označeni glede na vrsto podatka.

Eden izmed standardnih formatov, primernih za označevanje vsebine, je XML,[2] ki je v leksikografiji po svetu zelo razširjen. Smiselno ga je uporabiti za večnivojsko hierarhično strukturirane podatkovne zbirke (drevesna struktura), omogoča povezovanje s sklici, zahtevno iskanje in različne obdelave podatkov. Oznake elementov je mogoče določiti skoraj poljubno. Privzeto je uporabljeno unikodno kodiranje znakov, kar skupaj z ustrezno unikodno pisavo predstavlja dovolj možnosti za uporabo najrazličnejših znakov. Z datotekami XML delujejo številni leksikografski programi – na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU uporabljamo program iLEX,[3] med bolj znanimi so še ABBYY Lingvo Content,[4] IDM DPS,[5] TshwaneLex,[6] v slovenskem okolju pa je nastala Termania.[7] Ker so datoteke XML pravzaprav navadne besedilne datoteke, so prenosljive med različnimi programi in operacijskimi sistemi, kar je dolgoročno velikega pomena, saj lahko slovarsko podatkovno zbirko ne glede na uporabljeni program za izdelavo pozneje uporabimo v katerem koli programu, ki zna brati navadne besedilne datoteke.

Za celostno uporabnost datotek XML so bili znotraj družine XML razviti različni jeziki.[8] Jezik XSLT je namenjen preoblikovanju dokumenta XML v druge formate, kar lahko pomeni spreminjanje samega dokumenta XML ali pa pretvorbo v formate za prikaz na zaslonu ali za tisk. HTML in XHTML sta formata, ki se

[2] http://www.w3.org/standards/xml/
[3] http://www.emp.dk/
[4] http://www.abbyy.com/lingvo_content/
[5] http://www.idm.fr/products/dictionary_writing_system_dps/27/
[6] http://tshwanedje.com/tshwanelex/
[7] http://www.termania.net/
[8] http://www.w3schools.com/.

uporabljata za prikaz na zaslonu (najbolj tipičen primer so spletne strani), XSL:FO pa je primeren za pretvorbo podatkov v PDF, torej za tisk. Tovrstne pretvorbe so za človeško uporabo skoraj nujne, ker je XML označevalni jezik in ni lepo berljiv, ljudje pa želimo na zaslonu ali na papirju videti oblikovano besedilo. Za iskanje podatkov se uporablja XQuery, za navigacijo XPath, strukturo datotek XML pa določajo sheme različnih formatov – DTD, XML Schema[9] (.xsd) in RELAX NG (.rng). Drugačen tip sheme predstavlja ISO Schematron (.sch), ki preverja, ali je vsebina nekega elementa dovoljena glede na vsebino drugega elementa.

2 Slovarska struktura

V času izdelave koncepta slovarja je ključnega pomena predvideti tako strukturo slovarja, ki bo veljala za vse sestavke tega slovarja. Pri tem seveda ni nujno, da se vsi možni sestavni deli slovarskih sestavkov pojavljajo v vseh slovarskih sestavkih. Naloga izdelave takšne strukture je še posebej zahtevna pri specializiranih slovarjih, v katerih želimo prikazati veliko različnih podatkov ali pa je sama narava jezikovnih podatkov zelo raznovrstna, pa tudi sicer je načrtovanje ustrezne strukture eden ključnih korakov pred začetkom redakcijskih del.

Formalno strukturo slovarske podatkovne zbirke v XML-formatu opisuje shema. Ta določa, kateri elementi so v slovarski zbirki dovoljeni, kakšna so hierarhična razmerja med njimi in kakšen je njihov vrstni red, kakšne so možnosti njihovega kombiniranja oz. izključevanja in kolikokrat se določen element lahko ponovi, kadar želimo navesti več zaporednih enakih elementov. Shema ne nazadnje določa tudi, kakšna sme biti formalna vsebina elementov: ali je dovoljeno vsakršno besedilo ali obstaja omejitev na seznam možnih izbir (spustni meni) ali omejitev po dolžini vsebine, omejitev samo na številke, možnost vsebovanja atributov itd. Shema je torej rezultat nekaterih strukturnih odločitev, zajetih v slovarskem konceptu, saj lahko v zvezi z leksikografsko vsebino pomaga le pri nekaterih tehničnih zahtevah, nikakor pa ne more preprečiti vsebinskih neustreznosti, ki niso skladne s konceptualnimi napotki.

Obstaja več standardnih formatov shem za XML-dokumente; danes verjetno najbolj razširjena sta formata DTD in XML-shema.[10] Kljub nekoliko večji uporabnosti formata XML-shema v primerjavi z DTD-jem je uporaba enega ali drugega formata sheme običajno določena z leksikografskim programom, saj uporabe različnih formatov shem nekateri programi ne omogočajo.[11]

Ena od osnovnih nalog programa za leksikografsko delo je, da skrbi za skladnost slovarskih sestavkov s shemo in opozarja na nepravilnosti v strukturi in formalni vsebini. Tako imajo leksikografi orodje, ki jih usmerja k enotnosti strukture vseh slovarskih

[9] V besedilu je uporabljen poslovenjen zapis XML-shema.

[10] *XML-shema* (datoteka vrste .xsd) in *shema* nista sinonimna izraza, saj je *XML-shema* (dejansko: *XML Schema*) tako kot *DTD* le eden od formatov shem, ki opisujejo strukturo vsebine datoteke XML (Hunter idr. 2007: 145). Več o XML-shemi Thompson idr. 2004.

[11] IDM DPS uporablja DTD, iLEX in Termania uporabljata XML-shemo, Tshwanelex ima svoj interni DTD.

sestavkov, kar je zlasti koristno pri večjih projektih, kjer sodeluje večje število leksikografov. Pri tem je treba opozoriti, da se izdelava sheme v praksi ne zaključi nujno pred začetkom izdelave slovarja, temveč so manjše spremembe in izpopolnitve možne tudi med samim procesom izdelave slovarja. To se potrjuje tudi pri slovarskih projektih, ki potekajo na Inštitutu za slovenski jezik ZRC SAZU, zato je v času poskusne izdelave slovarskih sestavkov smiselno, da so sestavljavci slovarja in računalniški sodelavci pozorni na morebitne neustreznosti slovarske strukture in posledično sheme.

Pri pripravi shem za slovarje, ki nastajajo na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU, se je pokazalo, da je smiselno upoštevati več vidikov, ki vplivajo na izdelavo sheme, in sicer:

- leksikografski oz. vsebinski vidik,
- praktični vidik,
- tehnični vidik.

2.1 Leksikografski oz. vsebinski vidik

Leksikografski oz. vsebinski vidik ne pomeni nič drugega kot to, da naj bo v idealnem primeru slovarska struktura taka, kot si jo leksikograf predstavlja glede na vsebino, ki jo želi ustvariti. V praksi to pomeni najprej določitev in poimenovanje sestavnih delov slovarja (elementov sheme), nato pa določitev razmerij med temi elementi in dovoljene vsebine elementov. Dodatno se je mogoče odločati o združevanju več elementov v nadrejeni element, če je táko leksikografsko videnje strukture in če je to koristno zaradi razporeditve podatkov ali preglednosti. Tak primer je lahko npr. element zaglavje, ki vsebuje tiste podelemente, ki jih leksikograf vidi v zaglavju, ne pa, recimo, pri pomenu.

Kot primer preprostega dokumenta XML lahko prikažemo kazalčni slovarski sestavek *bodočnost*, katerega vsebina je povzeta po delovnem gradivu za Slovar sinonimov slovenskega jezika:[12]

```
<slovarski_sestavek>
   <iztočnični_del>
      <iztočnica>bodóčnost</iztočnica>
      <obrazilo>-i</obrazilo>
      <besednovrstna_oznaka>
         <samostalnik>ž</samostalnik>
      </besednovrstna_oznaka>
   </iztočnični_del>
   <kazalčni_del>
      <ciljna_dominanta>prihodnost</ciljna_dominanta>
   </kazalčni_del>
</slovarski_sestavek>
```

V tem primeru je uporabljenih osem različnih elementov, vsi so vsebinsko jasni in nedvoumni, njihova medsebojna razmerja so jasna in logična. Vsak element se

[12] Zamik v desno je uporabljen za označitev podrejenih elementov.

začne z začetno oznako, npr. <slovarski_sestavek>, in konča s končno oznako, npr. </slovarski_sestavek>, med tema oznakama pa imamo besedilo ali druge hierarhično podrejene elemente.

2.2 Praktični vidik

Pri slovarju s kompleksno mikrostrukturo število različnih elementov lahko hitro doseže trimestno število. Poleg tega se lahko nekateri elementi pojavljajo na različnih mestih, kar je po eni strani smiselno, po drugi strani pa so razmerja in vrstni red določenih elementov lahko na različnih mestih ravno dovolj drugačna, da leksikografa spominsko preobremenijo. Prav zato je smiselno upoštevati praktični vidik slovarske sheme, saj mora biti ta ne le logična, temveč tudi obvladljiva in mora čim bolj naravno sovpadati z leksikografskim procesom. Vprašanje je namreč, ali bodo leksikografi lahko dovolj obvladovali kompleksno strukturo, ne da bi jih ta ovirala pri njihovem delu, prav tako pomembno pa je, da bodo podatki urejeni tako, da bodo tudi končni uporabniki slovarja zmogli podatke razumeti.

Hierarhično globoka struktura sheme oz. slovarja ima prednost pred manj hierarhizirano v tem, da je več podatkov mogoče obravnavati združeno v nadrejenem elementu, kar pripomore k poenostavljenemu iskanju in obravnavi podatkov, prav tako so podatki lahko strukturirani skladno z leksikografovim razumevanjem strukture slovarja. Po drugi strani pretirano razvejana struktura otežuje orientacijo znotraj slovarskega sestavka (posledično lahko podaljša čas sestavljanja sestavkov), pozornost preusmerja k skrbi za pravilnost hierarhije namesto k vsebini, zato je na svoj način zahtevna za delo.

2.3 Tehnični vidik

S tehničnega vidika niso pomembna le razmerja med elementi, ampak tudi druge možnosti, ki jih želimo omogočiti. Tu gre zlasti za sklice (če želimo, da bo elektronska različica klikljiva, oz. če želimo v redakcijskem procesu zagotoviti enotnost nekaterih vsebin), razmerje med mešanimi vsebinami proti strukturiranosti ipd. Pomemben dejavnik je tudi sam leksikografski program – ta naj načeloma ne bi prav dosti vplival na izdelavo sheme, vendar se v praksi izkaže, da ni čisto tako. Pomemben vpliv imata lahko uporabniški vmesnik programa in način iskanja podatkov, saj uporabniški vmesnik dejansko vpliva na odločitev o tem, ali je smiselno ponavljajoče se elemente združevati v en nadelement ali ne, ker si lahko v nekem trenutku želimo na zaslonu skrčiti oz. skriti določene hierarhično povezane elemente, da imamo boljši pregled nad drugo vsebino, ki se ji želimo posvetiti. Programi se preglednosti vmesnika lotevajo različno, zato brez upoštevanja tega vidika ne gre.

Za iskanje podatkov in njihovo preurejanje, pretvorbo ali drugačno nadaljnjo uporabo je pomembna hierarhična razporeditev elementov, pojavitev istega elementa na različnih mestih pa lahko zaradi različnih kombinacij pojavitev naredi iskanje relevantnih podatkov zapleteno, obenem pa se poveča nevarnost dobivanja neustreznih zadetkov iskanja. Prav iskanje po že obstoječih vsebinah je pomemben vidik pri leksikografskem delu, saj daje vpogled v že preverjene slovarske rešitve in pomaga pri vzdrževanju enotnosti skozi slovar ter je tako koristno dopolnilo k

priročniku z napotki za sestavo slovarja. Potem ko je slovar dokončan, je slovarsko bazo mogoče uporabiti tudi za jezikoslovne raziskave, zato je način ureditve podatkov v njej lahko pomemben tudi s tega vidika.

Občasno so potrebni elementi z mešano vsebino,[13] tj. elementi, ki lahko vsebujejo pomešano besedilo in elemente s svojo vsebino. To je uporabno zlasti takrat, ko želimo v besedilo dodati drugače oblikovano besedilo, npr. nadpisane in podpisane številke (m^3, CO_2), sicer pa je to pogosta rešitev za etimološki razdelek, kjer so jeziki, primeri in navadno besedilo »pomešani« med seboj. S tehničnega vidika se na splošno priporoča previdnost pri uporabi mešanih vsebin, ker je nadaljnje procesiranje teh podatkov lahko bistveno težje kot sicer.

Kako skušamo pri pripravi XML-sheme navedene dejavnike (omenjeni so seveda le nekateri od možnih) upoštevati oz. uresničevati, hkrati pa njihovo udejanjanje čim bolj smiselno uravnotežiti glede na zahteve specifičnega slovarskega projekta, si v nadaljevanju oglejmo na primeru izdelave XML-sheme za Slovar sinonimov slovenskega jezika, ki nastaja na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU.

3 Izdelava XML-sheme za Slovar sinonimov slovenskega jezika[14]

Na oblikovanje XML-sheme za Slovar sinonimov slovenskega jezika je poleg konceptualizacije hierarhične podatkovne strukture kot projekcije koncepta slovarske strukture pri samih redaktorjih najbolj vplival tehnični oz. programski vidik, deloma pa tudi dileme v zvezi z vprašanjem, v kakšni obliki je slovarsko besedilo najbolj smiselno objaviti na spletu oz. v elektronski obliki. Glede na naravo podatkov, ki jih slovar prinaša – sinonimi so tisti element slovarske mikrostrukture, pri katerem že tradicionalno pričakujemo sklice na druge slovarske sestavke –, je bilo ključno vprašanje, kako oblikovati XML-shemo tako, da bo ta za redaktorje čim bolj logična, intuitivno razumljiva in zapomnljiva, hkrati pa čim bolj ustrezna za učinkovito in preprosto vzpostavljanje predpostavljenih sklicev na različne mikrostrukturne elemente slovarske podatkovne zbirke.[15]

[13] Angl. *mixed content*.

[14] Slovarsko gradivo, ki ga navajamo v nadaljevanju, je delovno gradivo raziskovalne skupine, ki pripravlja Slovar sinonimov slovenskega jezika in jo sestavljajo Martin Ahlin, mag. Branka Lazar, Zvonka Praznik in dr. Jerica Snoj. Izhodiščna XML-shema je bila pripravljena v sodelovanju z dr. Jerico Snoj, vsi sodelavci skupine pa so predlagane rešitve preizkusili v praksi in s svojimi pripombami, predlogi in dopolnitvami sooblikovali njeno končno različico. Za dragocene pripombe k članku se avtorja zahvaljujeva dr. Jerici Snoj.

[15] Koncept slovarja je bil oblikovan v času, ko XML še ni bil vzpostavljen kot standardni format zapisa jezikovnih podatkovnih zbirk, zato je bilo pri pripravi XML-sheme med drugim treba upoštevati že vzpostavljene rešitve v nastajajoči slovarski podatkovni zbirki.

3.1 Mikrostruktura slovarskih sestavkov Slovarja sinonimov slovenskega jezika

Glede na zastavljeni koncept podatkovno zbirko Slovarja sinonimov slovenskega jezika v strukturnem smislu sestavljajo slovarski sestavki dveh tipov – polni oz. dominantni slovarski sestavki in kazalčni oz. nedominantni slovarski sestavki.[16] Dominantni slovarski sestavki so iz treh delov: iztočnični del vključuje podatke o iztočnični besedi, kot so v Slovarju slovenskega knjižnega jezika navedeni v zaglavju, tj. podatke o besedni vrsti iztočnice, njeni izgovorjavi, naglasu, oblikoslovnih lastnostih ter njenih morebitnih dvojnicah. V pomenskem delu – natančneje poimenovanem razlagalno-sinonimni del – slovar opozarja na (samo) tiste pomene geselske besede, v katerih iztočnica nastopa kot običajen, (najbolj) nevtralen, tj. dominantni leksem za izražanje konkretnega slovarskega pomena, poleg tega pa so pri vsakem od pomenov navedeni še drugi (eno- ali večbesedni) nedominantni sinonimi iztočničnega leksema. Slovar v tem razdelku vključuje tudi podatke o delnih sinonimih in antonimih iztočnice v konkretnih pomenih ter o razširjeni zamenljivosti[17] (za več podatkov prim. Ahlin idr. 2003). Poleg tega opozarja še na tiste večbesedne lekseme[18] – v katerih iztočnična beseda nastopa kot skladenjsko jedro besedne zveze –, ki so prav tako nevtralni, dominantni leksemi za izražanje določenega pomena, ob njih pa predstavlja še druge (eno- in večbesedne) nedominante sinonime omenjene besedne zveze. Tretji razdelek dominantnega slovarskega sestavka je po vsebini soroden kazalčnemu delu kazalčnih slovarskih sestavkov. Prinaša podatke o tem, v katerih pomenih iztočnica ni dominantni leksem za izražanje konkretnega pomena, ampak le eden od nedominantnih sinonimov k dominantnemu leksemu, obdelanemu v drugem dominantnem slovarskem sestavku. Primera dominantnih slovarskih sestavkov z iztočnicama *gozd* in *teči* prikazujeta sliki 1 in 2.

gòzd gôzda m

1. |z drevjem strnjeno porasel svet| *Nad vasjo se razprostira gozd*
◊ pokr. **boršt** ◊ pesn. **gaj** pokr. gor. **gošča** ◊ ekspr. **gozdek** ◊ ekspr. **gozdič** ◊ ekspr. **gozdiček** ◊ pokr. **hosta** ◊ neobč. **les** ◊ star. **lesovje** ◊ neobč. **log** ◊ neobč. **loza** ◊ pokr. **šuma**
♦ **gaj**↗ ♦ zastar. **gora** |v hribovitem svetu| ♦ gozd. **letvenik** |v katerem imajo drevesa debelino letev|

•
1. borov gozd ◊ **borovje** ◊ pokr. **borovec**
2. brestov gozd ◊ redk. **brestje**
3. brezov gozd ◊ **brezje** ◊ **brezovje** ◊ neobč. **brezova loza** ◊ neobč. **brezova lozica**

[16] Podrobnejši podatki o strukturi slovarja so dostopni v Ahlin idr. 2003.

[17] V Ahlin idr. 2003 je razdelek razširjena zamenljivost poimenovan podpomenke.

[18] Dominantni besednozvezni leksemi za izražanje konkretnega pomena so lahko umeščeni v okvir posameznih pomenov iztočničnega leksema, lahko pa se pojavljajo v okviru posebnega gnezda besednozveznih enot. Umestitev je odvisna od stopnje pomenske povezanosti skladenjsko jedrne besede dominantnega večbesednega leksema z navedenimi pomeni iztočničnega leksema.

4. bukov gozd ◊ **bukovje** ◊ knjiž. pog. **bukev** ◊ redk. **bukovec** ◊ redk. **bukovina** ◊ neobč. **bukov log** ◊ pokr. **bukova šuma**
5. cerov gozd ◊ **cerje** ◊ **cerovje**
6. gabrov gozd ◊ **gabrovje** ◊ redk. **gabrina** ◊ redk. **gabrje**
7. hrastov gozd ◊ **hrastje** ◊ **hrastovje** ◊ redk. **hraščina** ◊ star. **dobje** ◊ star. **dobov gozd** ◊ star. **dobovje** ◊ redk. **hrastina** ◊ redk. **hrastovina**
8. iglasti gozd ◊ gozd. **črni gozd** ◊ redk. **igličasti gozd** ◊ redk. **igličevje** ◊ redk. **igličje** ◊ redk. **iglovje**
9. javorov gozd ◊ **javorje** ◊ **javorovje**
10. jelov gozd ◊ **hojev gozd** ◊ **hojevje** ◊ **jelovje** ◊ redk. **jelkov gozd** ◊ redk. **jelovina**
11. jelšev gozd ◊ **jelševje** ◊ **jelšje**
12. kostanjev gozd ◊ **kostanjevje**
13. macesnov gozd ◊ **macesnovje** ◊ redk. **macesenje**
14. mladi gozd ◊ neobč. **mladje** ◊ pokr. vzh. **mladoles** ◊ pokr. **mladovje** ◊ redk. **podmladek** ◊ neobč. **pomladek**
15. smrekov gozd ◊ **smrečje** ◊ **smrekovje** ◊ redk. **smrečevje** ◊ redk. **smrečina** ◊ redk. **smrečnati gozd** ◊ redk. **smrekovec**

{pp:° drvnik↗, listnik, steljnik
° iglasti gozd↗, listnati gozd, mešani gozd}

2. |drevje, ki raste strnjeno skupaj| *Gozd zarašča pašnik*
◊ pokr. **hosta** ◊ neobč. **les** ◊ star. **lesovje**
♦ **gozdičevje** |nizko drevje|

GL. ŠE veliko (nedol. količ. štev. *gozd dimnikov*)

Slika 1: Dominantni slovarski sestavek z iztočnico *gozd*

têči têčem nedov.
1. *kam*, *kje* |premikati se s hitrejšimi koraki tako, da sta v določenem trenutku obe nogi odmaknjeni od podlage| *teči proti cilju*; *teči po stopnicah*
◊ ekspr. **cvirnati jo** ◊ ekspr. **cvreti jo** ◊ ekspr. **leteti** ◊ ekspr. **sprintati** ◊ ekspr. **ucvirati jo** ♦ ekspr. **brusiti pete** |hitro| ♦ ekspr. **dirjati↗**

2. *kam*, *kje* |premikati se tako, kot je značilno za tekočino| *Voda teče med skalami*
◊ ekspr. **brzeti** ◊ neobč. **strujati** ◊ neobč. **strujiti** ◊ neobč. **točiti se**

{pp:° 'počasi' cediti se, cezeti, cizeti, cureti, curljati, lesti, mezeti, polzeti, solzeti
° 'slišno, oddajajoč značilen zvok' čapljariti, čofotati, hahljati, vrvrati
° 'silovito' dreti, hudouriti, liti, ulivati se}

3. *brezos.*, *kje* |izraža, da se na mestu, imenovanem v določilu, v manjši meri pojavlja tekočina| *Iz rane teče*
◊ **cediti se** ◊ **cezeti** ◊ **cizeti** ◊ **cureti** ◊ **curljati** ◊ **mezeti**

4. |neprenehoma, brez prekinitve se nadaljevati, razvijati| *Dela tečejo*
◊ publ. **biti v teku** ◊ neobč. **potekati**

GL. ŠE bežati (teči *pred kom*), delovati (*Ura ne teče, če ni navita*), hiteti (*Nimam časa, tečem kupit kruh*), izlivati se (*Sava teče v Donavo*), iztekati (*Iz rane teče gnoj*), minevati (*Dnevi hitro tečejo*), potekati (*Meja teče ob reki*), premikati se (*Jermen, speljan po kolesu, že teče* 'ne miruje več'), prihajati (*Denar teče v blagajno*), tek (teči = *gojiti tek*; *tekmovati v teku*), usipati se (*Pesek teče v posodo*)

Slika 2: Dominantni slovarski sestavek z iztočnico *teči*

Kazalčni slovarski sestavki vključujejo iztočnični del, ki je enak iztočničnemu delu dominantnih slovarskih sestavkov, v kazalčnem delu pa prinašajo sklice na ustrezne dele dominantnih slovarskih sestavkov, v katerih se pojavljajo zgolj kot neprednostni, nedominantni sinonimi k iztočničnemu leksemu kot dominantnemu leksemu za izražanje konkretnega pomena. Natančnejši oris strukture kazalčnega slovarskega sestavka z iztočnico *gozdič* predstavlja slika 3.

gozdič -íča m, **GL.** gozd 1

Slika 3: Kazalčni slovarski sestavek z iztočnico *gozdič*

3.2 Oblikovanje XML-sheme za slovar

Pri oblikovanju XML-sheme za Slovar sinonimov slovenskega jezika se je glede na sklicno naravo slovarja kot izhodiščno in temeljno pojavilo vprašanje, kako v tehničnem smislu zagotoviti, da bodo vse polnopomenske besede, ki so v slovarju navedene kot sinonimi oz. deli sinonimov, vključene v slovar kot iztočnice ali drugi mikrostrukturni elementi slovarskih sestavkov, predvsem pa, kako poskrbeti, da bodo ustrezni elementi kazalčnih in dominantnih slovarskih sestavkov s sklici pravilno povezani na različne elemente dominantnih slovarskih sestavkov. Programsko bi bilo npr. mogoče oba navedena problema reševati z (ročnim in avtomatskim) vzpostavljanjem sklicev – takšna rešitev bi bila npr. koristna z vidika zagotavljanja pravilnega sklicevanja na homonimne iztočnice oz. z njih –, vendar bi bil proces z redakcijskega vidika precej neučinkovit in časovno potraten, programsko pa razmeroma zapleten,[19] predvsem pa bi strukturo XML-sheme praktično v celoti določal tehnični vidik vzpostavljanja sklicev, kar je v popolnem nasprotju s konceptualizacijo slovarske strukture pri redaktorjih in naravo redakcijskega procesa pri pripravi konkretnega slovarja. Prav zato smo se odločili, da proces vključitve vseh ustreznih iztočnic v slovarsko podatkovno zbirko tehnično nadzorujemo s pomočjo drugih programskih mehanizmov, potrebne vsebinske sklice pa vzpostavimo v zaključni fazi priprave slovarja. V nadaljevanju skušamo orisati, kako smo razmišljali o dejavnikih, ki so odločilno vplivali na vzpostavitev nakazanih rešitev.

Eno od ključnih vprašanj, ki smo si jih v začetni fazi priprave XML-sheme za Slovar sinonimov slovenskega jezika zastavljali, je vprašanje, ali je kazalčne slovarske sestavke treba v XML-shemi (v prvi fazi dela) upoštevati oz. jih v slovarski podatkovni zbirki sploh redigirati.[20] Spraševali smo se namreč, ali ne bi bilo na podlagi podatkov v pomenskem delu dominantnih slovarskih sestavkov večine

[19] Rešitev bi npr. pomenila, da bi bilo treba vzpostaviti več sto tisoč sklicev, ki pravzaprav ne bi bili sklici v vsebinskem smislu, tj. sklici, ki bi uporabnika dejansko vodili na mesta v slovarski strukturi, na katerih bi dobil vse potrebne jezikovne podatke, ampak bi zgolj zagotavljali, da so v slovarski strukturi podatki vpisani na vsa potrebna mesta.

[20] Kazalčni slovarski sestavki so gotovo pomembni za potencialno knjižno izdajo slovarja, saj uporabnik iskanega podatka sicer ne bi našel, v elektronsko izdajo slovarja pa jih morda sploh ne bi bilo treba vključiti. Ustrezen iskalnik bi namreč iskalne zadetke lahko razvrščal in jih prikazoval tako, da potreba po sklicevanju morda ne bi niti nastala.

ustreznih podatkov izvoziti v kazalčne slovarske sestavke in v razdelek dopolnjevalne kazalke dominantnih slovarskih sestavkov v zaključni fazi priprave slovarja in jih ročno dopolniti, kjer avtomatski izvoz oz. generiranje podatkov nista mogoča.[21] Izkazalo pa se je – in to je eno od dejstev, ki je ključno vplivalo na odločitev, da sklice v okviru slovarske podatkovne zbirke vzpostavimo šele, ko bo sámo slovarsko besedilo praktično pripravljeno –, da med samim redakcijskim procesom dominantnih in kazalčnih slovarskih sestavkov ni mogoče vedno jasno razmejevati oz. da v določeni fazi priprave slovarja ni mogoče z gotovostjo napovedati, ali bo konkretna iztočnica izhodišče kazalčnega ali dominantnega slovarskega sestavka, saj se strukturni tip konkretnega slovarskega sestavka lahko določi šele ob redakciji veliko različnih iztočnic. Odločitev o tipu slovarskega sestavka, katerega izhodišče je posamezna iztočnica, se namreč vzpostavi na podlagi analize sinonimnih razmerij, ki jih vzpostavljajo vsi leksemi, ki izražajo pomene, ki jih je mogoče izraziti tudi s konkretno iztočnico. Navedeno dejstvo ima neprijetno posledico, da se tip in posledično mikrostruktura posameznega slovarskega sestavka lahko spremeni praktično v kateri koli fazi priprave slovarskega besedila, ob morebitnem vzpostavljanju sklicev že med samo redakcijo posameznih slovarskih sestavkov pa bi redaktorji morali ob vsaki spremembi preveriti tudi ustreznost že vzpostavljenih sklicev oz. vzpostaviti nove.

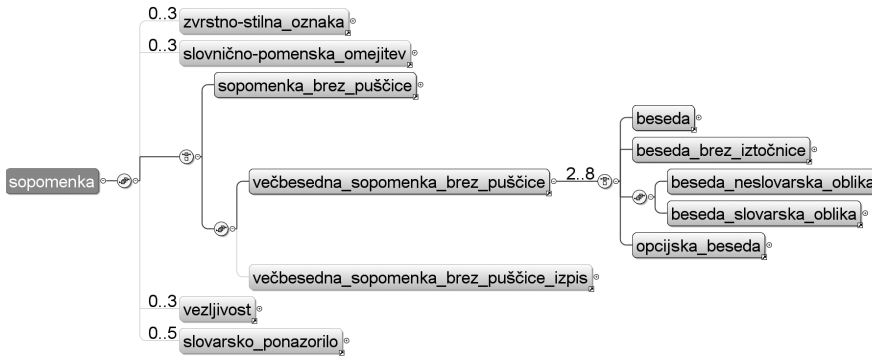
Tudi ko je bila odločitev o tem, da sklice vzpostavimo ob koncu procesa oblikovanja slovarskega besedila, že sprejeta, se je nekoliko še vedno postavljalo vprašanje, kako strukturirati elemente XML-sheme in koliko različnih elementov oz. podatkovnih tipov naj ta vsebuje, da bo redaktorski proces čim bolj učinkovit. Za samega redaktorja bi bilo po eni strani najbolj prikladno, če bi bilo elementov, v katere vpisujemo istovrstne ali sorodne informacije – ne nazadnje slovarsko strukturo v večini sestavljajo nizi (nedominantnih) sinonimov –, čim manj. Če bi namreč ugotovil, da mora mikrostrukturo slovarskega sestavka v celoti spremeniti, ker se je predpostavljeni strukturni tip slovarskega sestavka, katerega redakcijo pripravlja, spremenil, bi tako slovarsko besedilo popravil hitreje. Po drugi strani je treba upoštevati vidik vzpostavljanja sklicev. Če bi se odločili za malo različnih elementov sheme, bi se ti lahko sklicevali na več različnih mest v strukturi dominantnih slovarskih sestavkov, kar bi bilo z vidika prikaza možnih tarč sklicev v okviru vmesnika slovaropisnega programa in izbire ustreznega elementa v pogovornem oknu manj primerno, saj je več možnosti za napake, če je izbir veliko, poleg tega pa bi bilo v veliki meri onemogočeno avtomatsko vzpostavljanje sklicev. Manjše število podatkovnih tipov je morda manj ustrezna rešitev tudi z vidika iskanja po slovarski podatkovni zbirki, saj mora redaktor, da bi dobil ustrezne iskalne zadetke, oblikovati bolj kompleksen iskalni pogoj, pri čemer lahko znova prihaja do napak.

Druga možnost je, da pripravimo XML-shemo, v kateri je večina mikrostrukturnih elementov slovarja obravnavana kot poseben element sheme oz. kot samostojen podatkovni tip. Sklici s tovrstnih elementov bi seveda bili glede tarčnih mest bolj omejeni, torej za potrebe konkretnega slovarja s tehničnega vidika nekoliko

[21] Ročno bi bilo npr. treba vnesti nekatere podatke v iztočničnem delu kazalčnih slovarskih sestavkov, npr. izgovor besede, prav tako bi bilo treba ročno preveriti in popraviti vnos homonimnih iztočnic.

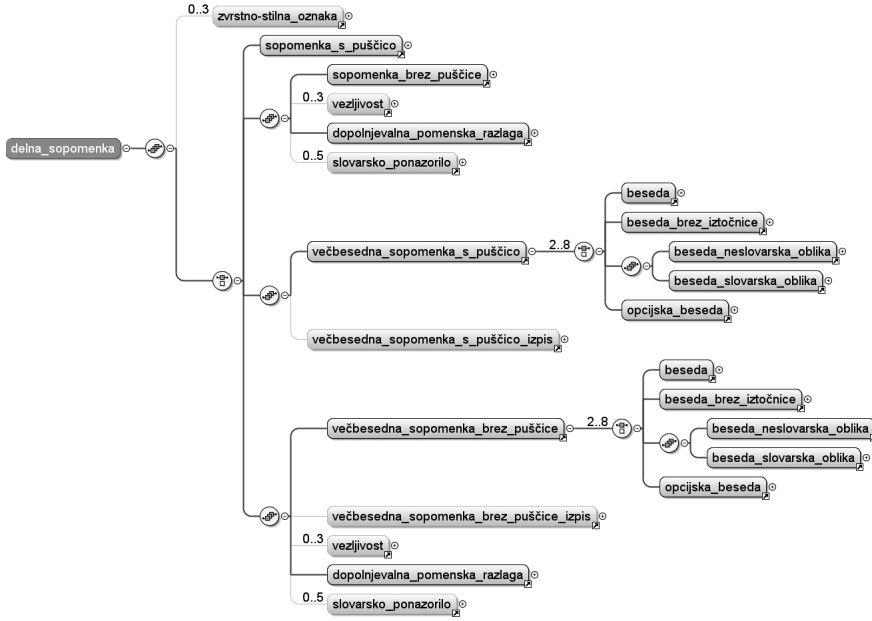
ustreznejši, če pa bi redaktor ob analizi gradiva ugotovil, da mora spremeniti tip slovarskega sestavka pri konkretni iztočnici, bi popravljanje mikrostrukture slovarskega sestavka zahtevalo precej več časa in napora.

Alternativnih možnosti je seveda še nekaj. Pri oblikovanju končne različice XML-sheme za Slovar sinonimov slovenskega jezika smo se odločili za kompromisno možnost, ki v nekoliko večji meri, kot je morda običajno sicer, upošteva tehnični vidik, tj. sklicevanje in želeni končni izpis podatkov. Različnih elementov sheme, v katere dejansko vpisujemo slovarske podatke o sinonimnih razmerjih, je razmeroma malo, vendar pa so ti umeščeni v jasno strukturirane različne nadelemente, ki so redaktorjem logični tudi glede na njihovo konceptualizacijo slovarske strukture, znotraj njih pa so omenjeni elementi na vseh ravneh slovarske strukture (pravi sinonimi, delni sinonimi, razširjena zamenljivost ...) hierarhizirani na enak način, zato XML-shema za redaktorje spominsko ni preveč obremenjujoča (prim. sliki 4 in 5, ki na primeru elementov (prava) sopomenka in delna sopomenka kažeta na številčno omejenost različnih shemskih elementov in njihovo enotno hierarhično strukturiranost[22]). Takšna obravnava se je zdela najbolj smiselna, saj smo, kot rečeno, skušali upoštevati tudi tehnične zahteve pri naknadni vzpostavitvi sklicev.



Slika 4: Struktura elementa sopomenka v XML-shemi Slovarja sinonimov slovenskega jezika

[22] Na slikah 4 in 5, ki prikazujeta segmente XML-sheme za slovar, okrepljena črta označuje obvezne elemente slovarske podatkovne zbirke, neokrepljena pa neobvezne. Simbol s tremi prekrivajočimi se sivimi pravokotniki označuje t. i. sekvenco, tj. obvezno zaporedje elementov v podatkovni zbirki, simbol, ki vključuje bel pravokotnik, pa označuje (obvezno) izbiro med naštetimi elementi. Sliko 4 bi torej lahko interpretirali nekako takole: element sopomenka, ki se lahko v slovarski strukturi pojavi od ničkrat do neskončnokrat, sestavljajo naslednji elementi: prvi neobvezni element je zvrstno-stilna_oznaka, drugi neobvezni element je slovnično-pomenska_omejitev, sledi natanko eden od obveznih elementov, tj. sopomenka_brez_puščice ali zaporedje elementov večbesedna_sopomenka_brez_puščice (slednji ima nekoliko kompleksnejšo strukturo) ter večbesedna_sopomenka_brez_puščice_izpis, poleg tega pa lahko element sopomenka sestavljata še neobvezna elementa vezljivost in slovarsko_ponazorilo v navedenem vrstnem redu. Vsi neobvezni elementi se lahko pojavijo večkrat, obvezni element pa natanko enkrat.



Slika 5: Struktura elementa delna sopomenka v XML-shemi Slovarja sinonimov slovenskega jezika

XML-shema za Slovar sinonimov slovenskega jezika glede na tip oz. mesto njihovega predpostavljenega sklicevanja predvideva, da so sinonimi umeščeni v shemske elemente dveh tipov (prim. sliko 6, ki prikazuje strukturo slovarskega sestavka *gozd*, na kateri so sklicevalni sinonimni oz. shemski elementi prvega tipa podčrtani z enojno črto, elementi drugega tipa pa so označeni s puščico in podčrtani s črtkano črto; izpis istega slovarskega sestavka v XML-formatu in z enakimi oznakami glede sklicev prinaša Priloga). Eno- in večbesedni[23] sinonimi, ki nimajo statusa nevtralnega, dominantnega leksema za izražanje konkretnega pomena, so v dominantnem slovarskem sestavku obravnavani v okviru štirih elementov sheme (tj. sopomenka_brez_puščice, beseda, beseda_slovarska_oblika, protipomenka), pri čemer bo v nadaljnjih fazah dela tem elementom avtomatsko dodan poseben atribut, ID-številka. Atribut ID bomo v zaključni fazi priprave slovarja skupaj z vsebino nekaterih drugih elementov slovarskega sestavka, zlasti neonaglašene iztočnice, avtomatsko izvozili v ustrezne, zlasti sklicevalne elemente kazalčnih in dominantnih slovarskih sestavkov,[24] na podlagi pripisanih atributov ID pa bodo vzvratno avtomatsko

[23] Koncept slovarja predvideva, da se kot samostojne iztočnice slovarskih sestavkov pojavijo tudi posamezne polnopomenske besede večbesednih sinonimov.

[24] Strukturni del dominantnega slovarskega sestavka, v katerem navajamo elemente z enakim tipom sklicevanja kot pri kazalčnih slovarskih sestavkih, so, kot smo že omenili, dopolnjevalne kazalke.

vzpostavljeni tudi sklici na (nad)elemente slovarske strukture dominantnih slovarskih sestavkov (npr. na posamezne pomene), v katerih so se kot nedominantni sinonimi pojavili.

Eno- ali večbesedni leksemi, ki so v katerem od svojih pomenov običajni, dominantni sinonim za izražanje konkretnega pomena (posledično so torej obravnavani tudi kot iztočnica v okviru dominantnega slovarskega sestavka), v konkretnem dominantnem slovarskem sestavku pa se pojavljajo v okviru dveh dopolnilnih razdelkov za izražanje sinonimnih razmerij, tj. med delnimi sopomenkami oz. v okviru elementa razširjena zamenljivost, so vpisani v dva druga shemska elementa (sopomenka_s_puščico, večbesedna_sopomenka_s_puščico). Ta se sklicujeta na različne elemente v pomenskem delu drugih dominantnih slovarskih sestavkov. Ker je z vidika vzpostavljenega slovarskega koncepta predstavitve sinonimnih razmerij med leksemi v slovenščini večina iztočnic enopomenskih in ker besedne zveze kot dominantni leksemi v slovarju večinoma ne nastopajo več kot enkrat, bo mogoče tudi sklice z omenjenih shemskih elementov na ustrezne elemente dominantnih slovarskih sestavkov večinoma vzpostaviti avtomatsko, ker bo imel iskalni izraz, ki išče tarčo sklica, le en ustrezen zadetek. Preostale sklice bomo povezali ročno. Ocenjujemo, da je tovrstnih elementov nekaj tisoč.

gòzd gôzda m

1. |z drevjem strnjeno porasel svet| *Nad vasjo se razprostira gozd*
◊ pokr. **boršt** ◊ pesn. **gaj** ◊ pokr. gor. **gošča** ◊ ekspr. **gozdek** ◊ ekspr. **gozdič** ◊ ekspr. **gozdiček** ◊ pokr. **hosta** ◊ neobč. **les** ◊ star. **lesovje** ◊ neobč. **log** ◊ neobč. **loza** ◊ pokr. **šuma**
♦ **gaj↗** ♦ zastar. **gora** |v hribovitem svetu| ♦ gozd. **letvenik** |v katerem imajo drevesa debelino letev|

•

1. borov gozd ◊ **borovje** ◊ pokr. **borovec**
2. brestov gozd ◊ redk. **brestje**
3. brezov gozd **brezje** ◊ **brezovje** ◊ neobč. **brezova loza** ◊ neobč. **brezova lozica**
4. bukov gozd ◊ **bukovje** ◊ knjiž. pog. **bukev** ◊ redk. **bukovec** ◊ redk. **bukovina** ◊ neobč. **bukov log** ◊ pokr. **bukova šuma**
5. cerov gozd ◊ **cerje** ◊ **cerovje**
6. gabrov gozd ◊ **gabrovje** ◊ redk. **gabrina** ◊ redk. **gabrje**
7. hrastov gozd ◊ **hrastje** ◊ **hrastovje** ◊ redk. **hraščina** ◊ star. **dobje** ◊ star. **dobov gozd** ◊ star. **dobovje** ◊ redk. **hrastina** ◊ redk. **hrastovina**
8. iglasti gozd ◊ gozd. **črni gozd** ◊ redk. **igličasti gozd** ◊ redk. **igličevje** ◊ redk. **igličje** ◊ redk. **iglovje**
9. javorov gozd ◊ **javorje** ◊ **javorovje**
10. jelov gozd ◊ **hojev gozd** ◊ **hojevje** ◊ **jelovje** ◊ redk. **jelkov gozd** ◊ redk. **jelovina**
11. jelšev gozd ◊ **jelševje** ◊ **jelšje**
12. kostanjev gozd ◊ **kostanjevje**
13. macesnov gozd ◊ **macesnovje** ◊ redk. **macesenje**
14. mladi gozd ◊ neobč. **mladje** ◊ pokr. vzh. **mladoles** ◊ pokr. **mladovje** ◊ redk. **podmladek** ◊ neobč. **pomladek**
15. smrekov gozd ◊ **smrečje** ◊ **smrekovje** ◊ redk. **smrečevje** ◊ redk. **smrečina** ◊ redk. **smrečnati gozd** ◊ redk. **smrekovec**

{pp:° drvnik↗, listnik, steljnik
° iglasti gozd↗, listnati gozd, mešani gozd}

2. |drevje, ki raste strnjeno skupaj| *Gozd zarašča pašnik*
◊ pokr. **hosta** ◊ neobč. **les** ◊ star. **lesovje**
♦ **gozdičevje** |nizko drevje|

GL. ŠE veliko (nedol. količ. štev. gozd dimnikov)

Slika 6: Vzorčni dominantni slovarski sestavek z oznakami, ki nakazujejo elemente enakega tipa z vidika sklicevanja

Ker Slovar sinonimov slovenskega jezika nastaja že nekaj let, je bila približno polovica predvidenih slovarskih sestavkov oblikovana z nespecializiranimi računalniškimi orodji. Na podlagi opisane XML-sheme bodo zato že obstoječi slovarski podatki pretvorjeni v XML-format, nato pa uvoženi v program iLEX, s pomočjo katerega že poteka redakcija novega gradiva. Jasna in pregledna hierarhična strukturiranost podatkovne zbirke, kot jo omogoča standardni XML-format, ob uporabi opisane XML-sheme in zmogljivih programskih orodij redaktorjem omogoča dober nadzor nad logično strukturo slovarske podatkovne zbirke, s čimer bo tudi nastajajoča slovarska zbirka bolj konsistentna in pregledno označena, za uporabnika pa posledično bolj relevantna in uporabna.

4 Zaključek

Tehnološki razvoj na področju računalništva je pomembno vplival tudi na sodobno leksikografijo – metodološko bi namreč sodobno leksikografsko delo lahko opredelili kot računalniško podprto. Kot standardni format za zapis slovarskih podatkovnih zbirk in tudi mnogih drugih jezikovnih virov se je v zadnjem času uveljavil XML, ki omogoča hierarhično strukturiranje podatkov, podprt z ustrezno shemo in ob uporabi sodobnih računalniških orodij pa redaktorjem omogoča, da odkrivajo nepravilnosti v strukturi in formalni vsebini podatkovne zbirke in posledično (ne)skladnosti pri slovarskih sestavkih. Pri oblikovanju sheme kot računalniške projekcije slovarskega koncepta je treba premišljeno upoštevati tri vidike, ki pomembno vplivajo na njeno ustreznost, tj. leksikografski oz. vsebinski vidik, praktični vidik ter tehnični vidik. Prav slednji se je pri vzpostavitvi XML-sheme za Slovar sinonimov slovenskega jezika zaradi sklicne narave slovarja z vidika sprejemanja konkretnih odločitev izkazal kot najodločilnejši za konkreten slovarski projekt.

Literatura[25]

ABBYY Lingvo Content ⟨http://www.abbyy.com/lingvo_content/⟩.

Ahlin idr. 2003 = Martin Ahlin idr., *Slovar sinonimov slovenskega jezika: splošna določila in opis zgradbe slovarskih sestavkov z vzorčno predstavitvijo*, Ljubljana: ZRC SAZU, Založba ZRC SAZU, 2003.

Clear 1987 = Jeremy Clear, Computing: Overview of the Role of Computing in Cobuild, v: *Looking Up: An Account of the COBUILD Project in Lexical Computing*, ur. John M. Sinclair, London – Glasgow: Collins ELT, 1987, 41–61.

Hunter idr. 2007 = David Hunter idr., *Beginning XML*, Indianapolis: Wiley Publishing, 2007.

IDM DPS ⟨http://www.idm.fr/products/dictionary_writing_system_dps/27/⟩.

iLEX ⟨http://www.emp.dk/⟩.

Sinclair 1987 = John M. Sinclair (ur.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, London – Glasgow: Collins ELT, 1987.

Standard XML ⟨http://www.w3.org/standards/xml/⟩.

Termania ⟨http://www.termania.net⟩.

Thompson idr. 2004 = Henry S. Thompson idr., *XML Schema Part 1: Structures: W3C Recommendation 28 October 2004* ⟨http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/⟩.

TshwaneLex ⟨http://tshwanedje.com/tshwanelex/⟩.

w3schools.com ⟨http://www.w3schools.com/⟩.

Priloga

```
<hom>
  <slovarski_sestavek>
    <iztočnični_del>
      <iztočnica>gòzd</iztočnica>
      <neonaglašena_iztočnica>gozd</neonaglašena_iztočnica>
      <obrazilo>gôzda</obrazilo>
      <besednovrstna_oznaka>
        <samostalnik>m</samostalnik>
      </besednovrstna_oznaka>
    </iztočnični_del>
    <razlagalno-sinonimni_del>
      <pomenska_enota>
        <razlaga>z drevjem strnjeno porasel svet</razlaga>
        <slovarsko_ponazorilo>Nad vasjo se razprostira gozd</slovarsko_ponazorilo>
        <sinonimni_niz>
          <sopomenka>
            <zvrstno-stilna_oznaka>pokr.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>boršt</sopomenka_brez_puščice>
          </sopomenka>
          <sopomenka>
            <zvrstno-stilna_oznaka>pesn.</zvrstno-stilna_oznaka>
```

[25] Vse navedene spletne strani so bile dostopne 9. 6. 2012.

```
            <sopomenka_brez_puščice>gaj</sopomenka_brez_puščice>
          </sopomenka>
          <sopomenka>
            <zvrstno-stilna_oznaka>pokr. gor.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>gošča</sopomenka_brez_puščice>
          </sopomenka>
          <sopomenka>
            <zvrstno-stilna_oznaka>ekspr.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>gozdek</sopomenka_brez_puščice>
          </sopomenka>
          <sopomenka>
            <zvrstno-stilna_oznaka>ekspr.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>gozdič</sopomenka_brez_puščice>
          </sopomenka>
[...]
          <sopomenka>
            <zvrstno-stilna_oznaka>pokr.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>šuma</sopomenka_brez_puščice>
          </sopomenka>
          <delna_sopomenka>
            <sopomenka_s_puščico>gaj</sopomenka_s_puščico>
          </delna_sopomenka>
          <delna_sopomenka>
            <zvrstno-stilna_oznaka>zastar.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>gora</sopomenka_brez_puščice>
            <dopolnjevalna_pomenska_razlaga>v hribovitem svetu</dopolnjevalna_pomenska_razlaga>
          </delna_sopomenka>
          <delna_sopomenka>
            <zvrstno-stilna_oznaka>gozd.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>letvenik</sopomenka_brez_puščice>
            <dopolnjevalna_pomenska_razlaga>v katerem imajo drevesa debelino letev
            </dopolnjevalna_pomenska_razlaga>
          </delna_sopomenka>
        </sinonimni_niz>
        <niz_besednozvezne_dominante>
          <besednozvezna_dominanta>
            <besedna_zveza>
              <beseda>borov</beseda>
              <beseda>gozd</beseda>
            </besedna_zveza>
            <besedna_zveza_izpis>borov gozd</besedna_zveza_izpis>
          </besednozvezna_dominanta>
          <sopomenka_BZ_dominante>
            <sopomenka_brez_puščice>borovje</sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
          <sopomenka_BZ_dominante>
            <zvrstno-stilna_oznaka>pokr.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>borovec</sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
        </niz_besednozvezne_dominante>
        <niz_besednozvezne_dominante>
          <besednozvezna_dominanta>
            <besedna_zveza>
              <beseda>brestov </beseda>
              <beseda>gozd</beseda>
              <besedna_zveza_izpis>brestov gozd</besedna_zveza_izpis>
```

```
            </besedna_zveza>
          </besednozvezna_dominanta>
          <sopomenka_BZ_dominante>
            <zvrstno-stilna_oznaka>redk.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>brestje</sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
        </niz_besednozvezne_dominante>
        <niz_besednozvezne_dominante>
          <besednozvezna_dominanta>
            <besedna_zveza>
              <beseda>brezov</beseda>
              <beseda>gozd</beseda>
              <besedna_zveza_izpis>brezov gozd</besedna_zveza_izpis>
            </besedna_zveza>
          </besednozvezna_dominanta>
          <sopomenka_BZ_dominante>
            <sopomenka_brez_puščice>brezje</sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
          <sopomenka_BZ_dominante>
            <sopomenka_brez_puščice>brezovje</sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
          <sopomenka_BZ_dominante>
            <zvrstno-stilna_oznaka>neobč.</zvrstno-stilna_oznaka>
            <večbesedna_sopomenka_brez_puščice>
              <beseda_neslovarska_oblika>brezova</beseda_neslovarska_oblika>
              <beseda_slovarska_oblika>brezov</beseda_slovarska_oblika>
              <beseda>loza</beseda>
            </večbesedna_sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
          <sopomenka_BZ_dominante>
            <zvrstno-stilna_oznaka>neobč.</zvrstno-stilna_oznaka>
            <večbesedna_sopomenka_brez_puščice>
              <beseda_neslovarska_oblika>brezova</beseda_neslovarska_oblika>
              <beseda_slovarska_oblika>brezov</beseda_slovarska_oblika>
              <beseda>lozica</beseda>
            </večbesedna_sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
        </niz_besednozvezne_dominante>
[...]
        <niz_besednozvezne_dominante>
          <besednozvezna_dominanta>
            <besedna_zveza>
              <beseda>smrekov</beseda>
              <beseda>gozd</beseda>
              <besedna_zveza_izpis>smrekov gozd</besedna_zveza_izpis>
            </besedna_zveza>
          </besednozvezna_dominanta>
          <sopomenka_BZ_dominante>
            <sopomenka_brez_puščice>smrečje</sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
          <sopomenka_BZ_dominante>
            <sopomenka_brez_puščice>smrekovje</sopomenka_brez_puščice>
          </sopomenka_BZ_dominante>
          <sopomenka_BZ_dominante>
            <zvrstno-stilna_oznaka>redk.</zvrstno-stilna_oznaka>
            <sopomenka_brez_puščice>smrečevje</sopomenka_brez_puščice>
```

```
      </sopomenka_BZ_dominante>
      <sopomenka_BZ_dominante>
        <zvrstno-stilna_oznaka>redk.</zvrstno-stilna_oznaka>
        <sopomenka_brez_puščice>smrečina</sopomenka_brez_puščice>
      </sopomenka_BZ_dominante>
      <sopomenka_BZ_dominante>
        <zvrstno-stilna_oznaka>redk.</zvrstno-stilna_oznaka>
        <večbesedna_sopomenka_brez_puščice>
          <beseda>smrečnati</beseda>
          <beseda>gozd</beseda>
        </večbesedna_sopomenka_brez_puščice>
      </sopomenka_BZ_dominante>
      <sopomenka_BZ_dominante>
        <zvrstno-stilna_oznaka>redk.</zvrstno-stilna_oznaka>
        <sopomenka_brez_puščice>smrekovec</sopomenka_brez_puščice>
      </sopomenka_BZ_dominante>
    </niz_besednozvezne_dominante>
    <razširjena_zamenljivost>
      <podpomenke>
        <snop>
          <člen_razširjene_zamenljivosti>
            <sopomenka_s_puščico>drvnik</sopomenka_s_puščico>
          </člen_razširjene_zamenljivosti>
          <člen_razširjene_zamenljivosti>
            <sopomenka_brez_puščice>listnik</sopomenka_brez_puščice>
          </člen_razširjene_zamenljivosti>
          <člen_razširjene_zamenljivosti>
            <sopomenka_brez_puščice>steljnik</sopomenka_brez_puščice>
          </člen_razširjene_zamenljivosti>
        </snop>
        <snop>
          <člen_razširjene_zamenljivosti>
            <večbesedna_sopomenka_s_puščico>
              <beseda>iglasti</beseda>
              <beseda>gozd</beseda>
            </večbesedna_sopomenka_s_puščico>
          </člen_razširjene_zamenljivosti>
          <člen_razširjene_zamenljivosti>
            <večbesedna_sopomenka_brez_puščice>
              <beseda>listnati</beseda>
              <beseda>gozd</beseda>
            </večbesedna_sopomenka_brez_puščice>
          </člen_razširjene_zamenljivosti>
          <člen_razširjene_zamenljivosti>
            <večbesedna_sopomenka_brez_puščice>
              <beseda>mešani</beseda>
              <beseda>gozd</beseda>
            </večbesedna_sopomenka_brez_puščice>
          </člen_razširjene_zamenljivosti>
        </snop>
      </podpomenke>
    </razširjena_zamenljivost>
  </pomenska_enota>
  <pomenska_enota>
    <razlaga>drevje, ki raste strnjeno skupaj</razlaga>
    <slovarsko_ponazorilo>Gozd zarašča pašnik</slovarsko_ponazorilo>
```

```
            <sinonimni_niz>
               <sopomenka>
                  <zvrstno-stilna_oznaka>pokr.</zvrstno-stilna_oznaka>
                  <sopomenka_brez_puščice>hosta</sopomenka_brez_puščice>
               </sopomenka>
               <sopomenka>
                  <zvrstno-stilna_oznaka>neobč.</zvrstno-stilna_oznaka>
                  <sopomenka_brez_puščice>les</sopomenka_brez_puščice>
               </sopomenka>
               <sopomenka>
                  <zvrstno-stilna_oznaka>star.</zvrstno-stilna_oznaka>
                  <sopomenka_brez_puščice>lesovje</sopomenka_brez_puščice>
               </sopomenka>
               <delna_sopomenka>
                  <sopomenka_brez_puščice>gozdičevje</sopomenka_brez_puščice>
                  <dopolnjevalna_pomenska_razlaga>nizko drevje</dopolnjevalna_pomenska_razlaga>
               </delna_sopomenka>
            </sinonimni_niz>
         </pomenska_enota>
         <dopolnjevalne_kazalke>
            <ciljna_dominanta>veliko</ciljna_dominanta>
            <razno>nedol. količ. štev. gozd dimnikov</razno>
         </dopolnjevalne_kazalke>
      </razlagalno-sinonimni_del>
   </slovarski_sestavek>
</hom>
```

Using XML Format in Lexicography: Creating an XML Schema for the *Dictionary of Slovenian Synonyms*

Summary

Because of technological progress, at the turn of the new millennium there were also key shifts in lexicography that had a significant impact on the methodology of lexicographic work, the conceptualization of lexicographic manuals, and, consequently, new methods for using them. Namely, for users dictionaries no longer merely represent manuals in book form, but also machine-readable databases appropriately tagged and structured. Because of its flexibility and universality, in recent years XML (eXtensible Markup Language) has become established as the standard format for records in dictionary and other linguistic databases. This format enables simple tagging of various hierarchically structured data and supports Unicode character encoding; because XML files are usually text files, they can be transferred between various programs and operating systems, and so they are also suitable for long-term data storage. A suitable schema that defines the formal structure of the database in XML format, and in part also the formal content of its elements, allows users control over the logical structure of the database, and so machine-readable dictionary databases are generally more consistent and their content is higher quality, and as a result they are also more relevant to the user. This article shows how to design a dictionary database or its XML Schema based on the design of the XML Schema for the *Dictionary of Slovenian Synonyms*.