10.35539/LTNC.2022.0047.S.R.298.311

Vārdnīcas digitālā struktūra

Sanda Rapa

Publikācija tapusi Valsts pētījuma programmas "Humanitāro zinātņu digitālie resursi" projektā "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (Nr. VPP-IZM-DH-2020/1-0001).

Atslēgvārdi: e-leksikogrāfija, elektroniska vārdnīca, mikrostruktūra, makrostruktūra

Katrai valodai ir savs (īpašs, no citām valodām atšķirīgs) digitālais nospiedums, kas balstās uz valodu atšķirībām, specifiskajām iezīmēm. Tāds ir arī flekšīvajai latviešu valodai, un tāpēc tās lingvistiskajai uzkrāšanai un eksponēšanai digitālajā vidē nepieciešama īpaša pieeja: proti, ņemot vērā latviešu valodas specifiskās iezīmes, latviešu vārdam leksikogrāfiskā avotā (jo īpaši digitālajā vidē) varētu būt jānorāda plašāks gramatiskais apraksts, jāņem vērā tā locījumu paradigma un diakritiskās zīmes utt. Saukdams šīs specifiskās iezīmes par "daudzslāņaino vārdnīcas arhitektūru" (Spohr 2011), to apliecina arī Deniss Spors (*Dennis Spohr*). Turklāt viņš piebilst: "[..] [I]r nepieciešams tāds leksikogrāfisks rīks, kas spēj tikt galā ne tikai ar dažādiem lietotāju tipiem un situācijām, bet arī tāds, kas nodrošina nepieciešamos mehānismus vārdnīcas satura individualizēšanai – individuālā lietotāja skatu pielāgošanai." Elektronisko vārdnīcu leksikogrāfiskā uzbūve sasaucas ar drukāto izdevumu leksikogrāfisko sistēmu, taču dažos aspektos būtiski atšķiras. Šo kopīgo pazīmju un atšķirību identificēšana un definēšana ir šā raksta mērķis.

Valodu resursus digitālajā vidē mēdz iedalīt trijās grupās:

- dati (piemēram, datubāzes, korpusi, tēzauri, vārdnīcas, analītiskie rīki utt.);
- rīki (piemēram, tulkotāji, runas atpazīšanas un pareizrakstības rīki, teksta redaktora rīki utt.);
- metadati, kas ļauj informāciju ievadīt, izgūt un uztvert (Bird, Simons 2003: 376 u. c.).

Runājot par šo valodas digitālās pasaules kopumu, parasti tiek lietots termins "valodas digitalizācija" (*language digitization*), ar to saprotot "valodas dokumentēšanu un aprakstīšanu, kur rezultāti ir paredzēti mašīnlasīšanai, ne cilvēku lietošanai" (Abney 2011: 1). Šajā rakstā tiek aplūkota neliela, bet būtiska valodas digitalizācijas daļa – elektronisku vārdnīcu un datubāzu sastādīšana, uz ko tiek attiecināts termins "elektroniskā leksikogrāfija" jeb "e-leksikogrāfija" (Logan 1991; Fuertes-Olivera, Bergenholtz 2013; Granger, Paquot 2012 u. c.) un kas ir gan rīku, gan metadatu pamatā.

Elektroniskajai leksikogrāfijai Latvijā ir pavisam nesena vēsture, taču pārdesmit gados ir paveikts daudz, un liela daļa no pašlaik pieejamajiem digitālajiem lingvistiskajiem leksikogrāfiskajiem avotiem ir LU Latviešu valodas institūta (LU LaVI) un LU Matemātikas un informātikas institūta (LU MII) sadarbības rezultāts: satura autors ir LU LaVI, savukārt elektroniskā apstrāde ir LU MII darba rezultāts.

Starp šādiem avotiem ir gan digitālās vārdnīcas, gan datubāzes, kas kopumā aptver pārsimts tūkstošus valodas vienību, kur katrai pievienots lingvistisks apraksts, piemēri un reģistrācijas dati. Balstoties uz šo pieredzi, rakstā tiek aplūkota latviešu valodas vienību digitalizācijas vēsture, specifika un nozīme no digitālo rīku satura veidotāju viedokļa.

Ieskats Latvijas e-leksikogrāfijā

Elektroniskās leksikogrāfijas pirmsākumi meklējami jau 20. gadsimta 50.–60. gados, kad sāka domāt par drukāto vārdnīcu pārveidošanu mašīnlasāmā formā – taču galvenokārt datu klasificēšanai, kārtošanai, kodēšanai, lai atvieglotu vārdnīcu rakstīšanu (Logan, 1991: 351), nevis otrādi – lai uzreiz veidotu elektroniskas vārdnīcas. Latvijā šo pieeju pazīst jau 70.–80. gados, taču vēl 1982. gadā, kad Igaunijā jau tapušas vairākas elektroniskas vārdnīcas, leksikogrāfe Liene Roze raksta: "Pagaidām elektroniskās iekārtas vēl nav piemērotas masveida lietošanai, jo ir dārgas un darbs ar tām ir sarežģīts." (Roze 1982: 112) Latvijā elektroniskas vārdnīcas sāk veidot 90. gadu beigās Matemātikas un informātikas institūta Mākslīgā intelekta laboratorijā (galvenokārt Andreja Spektora vadībā). Pirmās elektroniskās vārdnīcas top kā drukāto vārdnīcu elektronisks atspulgs – informācija viens pret vienu tiek ievadīta elektroniskā vidē.

LU Latviešu valodas institūtā (LU LaVI) izstrādāto vārdnīcu ceļš uz digitālo vidi sākas 2001. gadā, kad no drukātas vārdnīcas par elektronisku avotu kļūst LU LaVI (agrāk Latvijas Zinātņu akadēmijas Valodas un literatūras institūta) pētnieku sastādītā "Latviešu literārās valodas vārdnīca". Tagad tā digitālajā vidē iekļaujas kā statisks resurss – tā ir LU LaVI sagatavoto un apgādā "Zinātne" no 1972. līdz 1996. gadam izdoto astoņu sējumu digitāla publikācija, kas strukturēta mašīnlasāmā formātā (LLVV, https://llvv.tezaurs.lv). Vārdnīcas digitalizēšana ļāva piemēru un nozīmju sistēmu padarīt uzskatāmāku, novērst gan pamanītās kļūdas, gan drukātam izdevuma diktētos ierobežojumus (piemēram, saīsinājumus rakstīt pilnos vārdos). LLVV ir pagaidām apjomīgākais digitālais leksikogrāfiskais avots (tajā iekļauti vairāk nekā 64 000 šķirkļu), taču tā netiek mainīta un pielāgota mūsdienu valodas situācijai, paliekot kā sava veida statisks, nemainīgs liecinājums par 20. gadsimta 80.–90. gadu normēto valodu. Šāda vārdnīca ar laiku kļūs par vēsturisku liecību, kas noderēs arī valodas vēstures pētījumiem, taču, neatspoguļojot valodas attīstību, tā neseko lingvistiskās datubāzes principiem, no kuriem galvenais ir dinamika – ne tikai digitālā atjauninājuma, bet arī satura pilnveidošanas aspektā (Hunston 2002).

Tāpēc jau kopš 1997. gada LU Latviešu valodas institūtā tiek izstrādāts arī dinamiskās "Mūsdienu latviešu valodas vārdnīcas" (MLVV, https://mlvv.tezaurs.lv) saturs (2022. gadā vairāk nekā 45 000 šķirkļu). Tā ir balstīta uz pēdējo gadu desmitu latviešu valodas materiāla, kas iegūts ne tikai no drukātiem izdevumiem, bet arī no interneta un mūsdienu latviešu valodas korpusa. MLVV ir pilna tipa vārdnīca, kurā tiek reģistrēti visi mūsdienu latviešu valodā dzirdētie vārdi. Turklāt tajā arvien tiek iekļauti arvien jauni vārdi, nevairoties no jebkuras stilistiskas nokrāsas, un agrāk reģistrēto vārdu nozīmes tiek rediģētas, papildinātas atkarībā no

novērotā mūsdienu lietojuma, tādējādi atspoguļojot semantiskās pārmaiņas latviešu valodas pamatleksikā. Atšķirībā no LLVV "Mūsdienu latviešu valodas vārdnīca" aptver dažādus latviešu valodas slāņus, iekļaujot arī izplatītākos neliterārās leksikas vārdus, barbarismus u. c.

Par pagaidām dinamisku var saukt arī "Latviešu valodas vēsturisko vārdnīcu" (LVVV, https://tezaurs.lv/lvvv/), kuras izstrāde uzreiz digitālā formā tika aizsākta 2004. gadā, pamatojoties uz 2003. gadā izveidoto latviešu valodas seno tekstu korpusu (http://senie.korpuss.lv/), kas tiek papildināts vēl līdz šim brīdim). Kaut arī LVVV aptver tikai 16.–17. gadsimta latviešu rakstu avotu leksiku, tā joprojām tiek turpināta un tiek papildināta ar vārdiem un piemēriem no jaunatklātiem avotiem. Patlaban vārdnīcā ir nepilni divi tūkstoši šķirkļu, taču pēc laika šai vārdnīcai varētu būt statiska avota statuss – ja izdosies aptvert un digitalizēt visus 16.–17. gadsimta avotus, tā būs pabeigta un tās saturs netiks būtiski mainīts.

Par dinamiskiem uzskatāmi arī LU Latviešu valodas institūtā izstrādātie pūļpakalpes rīki: 2017. gadā atklātā "Tautas vietvārdu datubāze" (TVD, www.vietvardi.lv) un 2022. gadā radītā "Apvidvārdu datubāze" (AD, www.apvidvardi.lu.lv). Kaut arī to saturu nav radījuši valodnieki, tie tomēr uzskatāmi par leksikogrāfiskiem avotiem, jo tajos atbilstoši elektroniskās leksikogrāfijas principiem izstrādāti vienas valodas vienības atspoguļošanai nepieciešamie lauki. Tos var aizpildīt jebkurš interesents (arī bez valodnieciskām zināšanām), taču iesniegtie dati pēc pārbaudes izmantojami tālāk valodnieciskos pētījumos un leksikogrāfiskos avotos. Šo rīku rašanos stimulēja vispārējā digitalizācija un tās iespējamo ieguvumu apjaušana, sabiedrības digitālo prasmju uzlabošanās un leksikogrāfiskās digitalizācijas iespējas. Ja statiskās vārdnīcas daļēji turpina drukāto izdevumu principus, tad pūļpakalpes izstrāde liek meklēt arvien jaunus vārdu reģistrācijas un vizualizācijas aspektus. Tieši pūļpakalpes vietnes, kurās darbojas ne tikai valodnieki, bet galvenokārt jebkurš ieinteresēts interneta lietotājs, liek plānot pēc iespējas arvien ērtāku un pārskatāmāku vārda digitālo ekspozīciju.

Leksikogrāfiskās digitalizācijas principi (salīdzinājumā ar drukātiem izdevumiem)

Mūsdienu valodas digitalizācija lielā mērā balstās uz gadu desmitos un pat simtos izstrādātām gramatikas un leksikogrāfijas atziņām. Īpaši tas sakāms par elektroniskajām vārdnīcām – tajās tiek ievēroti galvenie vārdnīcas mikrostruktūras (elementu izkārtojuma šķirklī), makrostruktūras (šķirkļu izkārtojuma vārdnīcā) un megastruktūras (ārpus vārdnīcas korpusa publicētās papildinformācijas) principi (par vārdnīcas struktūrām sk. Jērāne 2014). Joprojām ir jāpiekrīt Roberta Vešlera (*Robert Weschler*) un Krisa Pitsa (*Chris Pitts*) 2000. gadā izteiktajai atziņai, ka "elektroniskās vārdnīcas vēl ir pamatos drukātas vārdnīcas, kas ievietotas mikročipā" (Weschler, Pitts 2000) – tām ir noteiktas ātrākas funkcijas, taču principi ir nemainīgi. Daudz vairāk digitālajā pasaulē mainās vārdnīcu struktūru svarīgums un samērs. Piemēram, makrostruktūra, ņemot vērā to, ka elektroniskā vārdnīcā vārdu var sameklēt, to ierakstot meklēšanas logā, nevis meklējot vārdnīcas šķirkļus kādā noteiktas sistēmas sarakstā, nav tik būtiska

un negrozāma. Arī megastruktūras apjoms un izvietojums nav reglamentēts. Tāpat būtiski atšķiras sinhronisku (mūsdienu vai kāda samērā neliela laikposma) vārdnīcu un diahronisku (vēsturisku vai reģionālu) vārdnīcu digitalizācijas principi. Sinhroniskajām vārdnīcām ir plašāks horizontālais (sinhroniskais) lauks: piemēram, LLVV un daļēji arī MLVV mikrostruktūru paplašina gramatiskā informācija, kas sniedz ziņas par vārda pamatformu (sk. 1. attēlu). Turpretim diahroniskajām vārdnīcām apjomīgs ir vertikālais (diahroniskais) lauks: piemēram, LVVV līdztekus mikrostruktūras pamatelementiem sniegtas ziņas par vārda attīstību laikā un vārda rakstību dažādos avotos (sk. 2. attēlu), bet TVD un AD par svarīgu papildkomponentu izvēlēta lokalizācija (vārda ģeoreferencēšana) (sk. 3. attēlu).

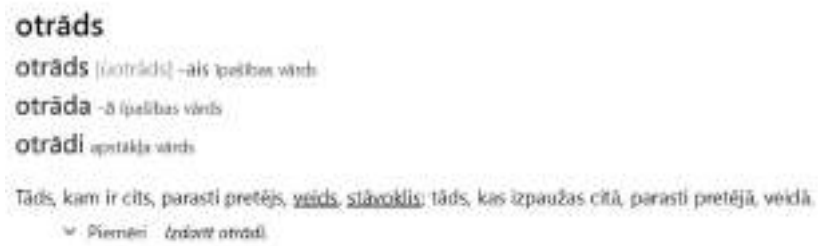
Mikrostruktūra

Elektroniskas vārdnīcas svarīgākais struktūras elements ir tās mikrostruktūra. Tāpat kā tradicionālajā leksikogrāfijā, arī leksikogrāfiskajā digitalizācijā to veido divi pamatelementi: vārds un tā nozīme. Pamatojoties uz vārda nozīmi un raksturu, tiek veidota elektroniskas vārdnīcas šķirkļa arhitektūra. Tai jābūt paredzamai, striktai un vienkāršai, ar strukturētu sintaksi (Abel 2012: 84), lai lietotājs tajā varētu viegli orientēties. Tāda ir arī LU LaVI elektronisko vārdnīcu struktūra – to pamatos veido vārds un tā nozīme, kas tiek ilustrēta ar piemēriem (vārda un nozīmes lauks ir obligāti aizpildāmās vienības, bez kurām leksisko vienību vārdnīcā nevar saglabāt). Vārds un nozīme, ko ilustrē piemēri, ir izvēlēti par galvenajiem elementiem arī pūļpakalpes rīkos – valodas datu uzkrāšanas vietnēs.

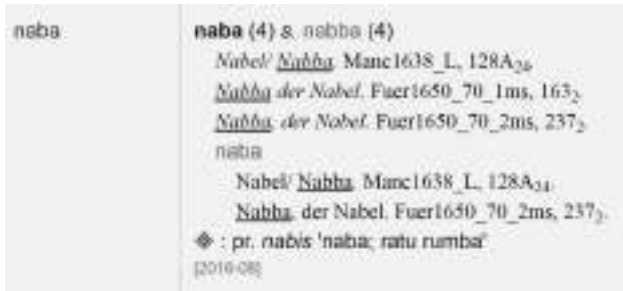
Vārds ir pamatelements valodas sistēmā, un vārdam jābūt galvenajam gan šķirkļa struktūras izkārtojumā, gan šķirkļu sarakstā. Tas tiek ievērots arī digitālajā leksikogrāfijā – vārds ir centrālais elements. Atšķirībā no tradicionālās leksikogrāfijas, kur vārds tiek tehniski izcelts (parasti trekninàts) šķirkļa struktūrā, digitālajā leksikogrāfijā vārds tiek nošķirts no šķirkļa struktūras – tas tiek eksponēts kā šķirkļa virsraksts. Elektroniskā leksikogrāfija atšķirībā no tradicionālās ļauj vārdu nemanāmi saskaldīt vēl sīkākās daļās – morfēmās un fonēmās, kuras var anotēt un tālāk izmantot klasificēšanā, strukturēšanā, analīzē, meklēšanā. Tas ir nenovērtējams ieguvums valodniekiem, kam agrāk šo visu funkciju nodrošināšanai vajadzēja izstrādāt atsevišķus leksikogrāfiskus avotus (piemēram, biežuma, inversās, sinonīmu utt. vārdnīcas).

Vārda lauka struktūra ir atšķirīga sinhroniskajā un diahroniskajā digitalizācijā. Sinhroniskajā digitalizācijā (mūsdienu vai kāda samērā neliela laikposma valodas digitālā atspoguļošanā) par šķirkļa vārdu tiek norādīta viena vārdforma, kurpretim diahroniskajā digitalizācijā (vēsturiskā valodas attīstības vai reģionālistikas atspoguļošanā) šķirkļa galvu var veidot vairāku vārdformu kopa (sal., piemēram, šķirkli *ābols* sinhroniskajā MLVV ar vienu pamatformu un diahroniskajā LVVV, kur šķirkļa galvu veido 36 vārdformas, kas reģistrētas dažādos avotos un dažādos laikposmos).

Nozīmi daži pētnieki uzskata par svarīgāko elektroniskas vārdnīcas elementu, jo tas ir galvenais, ko elektroniskā vārdnīcā meklē lietotājs (Lew 2010: 291). Tāpēc tai vārdnīcas sastādīšanā



1. attēls. "Mūsdienu latviešu valodas vārdnīcas" šķirklis.



2. attēls. "Latviešu valodas vēsturiskās vārdnīcas" šķirklis.

3. attēls. Apvidvārdu datubāzes šķirklis.

jāpievērš vislielākā vērība. Vārda nozīmes atspoguļošana elektroniskā vārdnīcā būtiski neatšķiras no tradicionālās leksikogrāfijas: tai tāpat jābūt precīzai, pietiekami īsai un noteiktai, pat neraugoties uz to, ka elektroniskā vidē nav nepieciešami tik lieli telpas un teksta ierobežojumi kā drukātos izdevumos. Nozīmi tāpat kā drukātos izdevumos var atspoguļot gan ar definīciju, gan ar ekvivalentu (sinonīmu), gan ar piemēru, taču elektroniskā leksikogrāfija ļauj bez šīm

tradicionālajām nozīmes ekspozīcijām izmantot arī citus risinājumus, piemēram, audiomateriālus, videomateriālus un neverbālus elementus (piemēram, attēlus, animācijas, neverbālas skaņas). Audiomateriāli ļauj daudz precīzāk uztvert valodas būtību, atspoguļo reālo valodu, savukārt neverbālie elementi palīdz izskaidrot grūti aprakstāmas parādības un priekšmetus (Lew 2010: 297ff.). Īpaši tas svarīgi reģionālajos valodas paveidos un izloksnēs, kur vārdu precīzā intonācijā var izrunāt tikai vietējais reģiona iedzīvotājs. Rēķinoties ar to, LU LaVI izstrādātajā "Apvidvārdu datubāzē" ir iekļauta iespēja vārdu arī ierunāt.

Digitalizācija ļauj viegli risināt dažus drukātā izdevumā grūti risināmus leksikogrāfiskus trūkumus nozīmju skaidrošanā. Piemēram, cirkulārās nozīmes (proti, nozīmes, kurās vārds skaidrots ar šķirkļa vārda daļām, piemēram, putekļsūcējs 'putekļu sūcējs'): aktīvu mijnorāžu izmantošana un vārdu daļu skaidrojuma ātra aplūkošana var būtiski atvieglot gan leksikogrāfa darbu, gan vārdnīcas lasītāja meklējumus. Tomēr no leksikogrāfiskā aspekta cirkulārās nozīmes arī leksikogrāfiskajā digitalizācijā nav vēlamas.

Nozīmes definēšanai gan drukātā, gan digitālā izdevumā vajadzētu izmantot proformas (Granger, Paquot 2012: 23): katrai semantiskajai kategorijai specifiskas definīcijas satura un struktūras veidnes, kurās viegli ievietot katra atšķirīgā vārda nozīmi vienotā formā, tādējādi uzlabojot vārdnīcas kopējo sistēmu. Elektroniskajā leksikogrāfijā proformas ir īpaši viegli pielāgojamas un izmantojamas gan šķirkļu veidošanā, gan to rediģēšanā, jo viegli atlasīt šādām veidnēm attiecīgās leksiskās ligzdas un pārskatīt nesakritības (Atkins, Rundell 2008: 123–124). Šādas proformas apzināti vai neapzināti ir atrodamas jebkurā vārdnīcā. Piemēram, LLVV dzīvnieka definēšanai lielākoties izmantota zoonīmiskā taksonomija apvienojumā ar tā uzvedības pazīmēm vai izmantošanas nolūku (piemēram, *kaķis* 'neliels plēsēju kārtas mājdzīvnieks, kas medī peles, žurkas, putnus'; *suns* 'suņu dzimtas mājdzīvnieks, kam ir saimnieciska vai dekoratīva nozīme'); Oksfordas vārdnīcā definīcijas veidnē iekļauts dzīvnieka lielums, apspalvojuma krāsa un taksonomija (turpat). Taču latviešu elektroniskajās vārdnīcās proformas ne vienmēr izmantotas – tas redzams citos šķirkļos, kur dzīvnieku definīcijā skaidrojuma struktūra ir citāda – nav pieminēta dzīvnieka zoonīmiskā taksonomija: piemēram, *zaķis* 'vidēji liels savvaļas dzīvnieks ar garām ausīm, pagarinātām pakaļkājām un īsu asti' (LLVV, MLVV).

Diahroniskajās vārdnīcās obligāti mikrostruktūras elementi ir hronotopi – laika un telpas marķieri (ne velti diahronisko lingvistisko avotu joma dažreiz tiek dēvēta par telpisko humanitāro zinātni (*spatial humanities*) (Won et al. 2018). Tā vietvārdu un apvidvārdu datubāzē par obligātu lauku noteikta lokalizācija, savukārt LVVV vēlams visu formu hronoloģisks atspoguļojums.

Piemēri, kaut arī valodas skaidrojošajās vārdnīcās nav obligāti aizpildāmi lauki, LU LaVI izstrādātajos elektroniskajos leksikogrāfiskajos avotos ir pievienoti visiem vārdiem, jo vārdnīcas lietotājam, kurš gan LLVV, gan MLVV uzskata par normējošiem avotiem, tie dažkārt ir pat izšķiroši un palīdz uztvert vēlamā vai nevēlamā lietojuma nianses. LU LaVI vārdnīcās vienmēr pievienota arī vārda gramatiskā informācija (vārdšķira, vārda dzimte, stilistiskā norāde, dažādi ierobežojoši vai izņēmuma gadījumi).

Makrostruktūra un megastruktūra

Vārdnīcas makrostruktūras un megastruktūras loma elektroniskajā vidē ir būtiski pavājinājusies. Pavisam nenozīmīga ir kļuvusi makrostruktūra – šķirkļu izkārtojums vārdnīcā. Elektroniskās leksikogrāfijas atbrīvošanos no alfabēta kārtības kā lielu ieguvumu un atslogu salīdzinājumā ar drukātajiem izdevumiem min visi e-leksikogrāfijas pētnieki (Granger, Paquot 2012; Lew 2010 u. c.). Īpaši to novērtēja un slavēja pirmo elektronisko vārdnīcu autori un lietotāji. Taču tagad, pēc vairākām aktīvām elektroniskās leksikogrāfijas desmitgadēm, lietotāji un vārdnīcu autori tomēr apjautuši alfabētiskas kārtības priekšrocības. Iespējams, šādu nepieciešamību radījusi valodas iekšējā paplašināšanās (daudzu atvasinājumu darināšana). Alfabētiskais saraksts ļauj viegli pārskatīt vārda apkaimi (visus vārda saknes atvasinājumus un līdzīgos vārdus). Nevienā no LU LaVI izstrādātajām elektroniskajām vārdnīcām šāda rādītāja nav. Lai kaut daļēji apmierinātu prasības pēc sistēmiska saraksta, MLVV tika izveidota sleja, kas rāda meklētā vārda apkaimi (piemēram, meklējot vārdu *skola*, vārdnīca dos ziņu, ka ir arī vārdi *skolasbērns*, *skolasnauda*, *skolēns*, *skolmeistars*, *skolniecisks*). Atvasinājumu reizēm ir tik daudz, ka tos visus nespēj parādīt pat apkaimes logs (tā MLVV *skolas* logā neredzēsim *skolnieks*, *skolot*, *skoloties* u. c.), tāpēc gan no valodnieku, gan lietotāju puses izskanējuši lūgumi pēc kopēja alfabētiska saraksta vai cita kopēja sistēmiska atspoguļojuma. Taču elektroniskajā leksikogrāfijā tas joprojām nav atrisināts – liela apjoma datiem tas, kas tradicionālajā leksikogrāfijā tik pierasts, elektroniskos avotos var kļūt par pārbaudījumu.

Arī megastruktūras lomas pavājināšanās digitālajās humanitārajās zinātnēs ir pamanīta (Burdick et al. 2012). To redz arī jebkurš lietotājs. Atšķirībā no drukāta izdevuma, kur līdz vārdnīcas tekstam var nokļūt caur vārdnīcas nosaukumu, autoru sarakstu un vārdnīcas struktūras aprakstu, elektroniskajā leksikogrāfijā vienīgais redzamais elements ir saīsināts vārdnīcas nosaukums (visbiežāk abreviatūra) un vārdnīcas elektroniskās vietnes nosaukums. Gandrīz nemanāms kļūst vārdnīcas autors vai sastādītājs. Digitālajās humanitārajās zinātnēs autors vairs nav autonoma persona, jo vārdnīcas sastādīšana nebeidzas ar šķirkļu sarakstīšanu. Tas, kas agrāk tika uzticēts izdevējiem (rediģēšana, izkārtojums, izplatīšana, pielāgošana lietotāju vajadzībām), tagad ir digitālo humanitāro zinātņu priekšmets (Burdick et. al. 2012: 83). Tāpēc arī vārdnīcas struktūras aprakstam digitālajā pasaulē tiek pievienots vārdnīcas elektroniskās ekspozīcijas skaidrojums.

Elektroniskās leksikogrāfijas ieguvumi

Kaut arī struktūru līmenī elektroniskā leksikogrāfija joprojām turpina tradicionālās leksikogrāfijas principus, tās ieguvumi gan no vārdnīcu lietotāju, gan sastādītāju puses tiek arvien vairāk novērtēti un arvien tiek paplašināti un pielāgoti. Apvienojot Silvianas Greindžeras (*Sylviane Granger*) (Granger, Paquot 2012: 2) atzinumus un LU LaVI pieredzi elektronisku vārdnīcu sastādīšanā, var izšķirt šādas elektroniskās leksikogrāfijas priekšrocības:

1. korpusa integrācija;
2. datu apjoma palielināšanās un kvalitātes uzlabošanās;
3. efektīva piekļuve;
4. ērta pielāgošana;
5. hibridizācija;
6. vienota sistēma;
7. lietotāja līdzdalība.

Korpusu integrācija (valodas korpusu datu iepludināšana elektronisko vārdnīcu šķirkļu izstrādē, nozīmes precizēšanā un piemēru bagātināšanā) kļūst arvien nozīmīgāka, un, palielinoties un uzlabojoties valodas resursiem, tā ir arvien efektīvāka. Elektroniskās leksikogrāfijas pētnieki uzskata, ka nākotnes vārdnīca pilnībā balstīsies uz kontekstualizāciju (vārdu nozīmes tiks izstrādātas, izvērtējot plašu to lietojumu kontekstu). Šķiet, nevienu mūsdienu vārdnīcu vairs nevarētu izveidot bez ātras un jaudīgas apjomīgu datu apstrādes, kas ļauj precīzi definēt vārdu nozīmes, savukārt lietotājiem dod iespēju iepazīt kontekstu. Šādas savstarpējas korpusa un vārdnīcas piemērs ir LVVV, kuras izstrādē tiek izmantots seno tekstu korpuss "Senie" un no kuras šķirkļa var nokļūt tieši pie izmantotās vārdformas konkrētā lietojuma vietā. Korpusu integrācija ļauj atrast un pievienot arvien vairāk piemēru un noteikt vārda lietojuma biežumu. Tas nozīmē, ka vienlaikus ar elektroniskas vārdnīcas izveidi jārūpējas arī par korpusa nodrošināšanu un paplašināšanu, jo, lai vārda nozīme būtu pēc iespējas precīzāka, tās izstrādes pamatā jābūt pēc iespējas lielākam korpusam.

Korpusu integrācija, valodas resursu uzkrāšana, uzglabāšana un apkopošana ir radījusi lielas iespējas arī leksikogrāfa darbā – tā rīcībā ir milzīgi valodu resursi, ko var izmantot leksikogrāfisko avotu izstrādes pamatā. Datu apjoma paplašināšanās paaugstina arī leksikogrāfisko avotu kvalitāti. Izmantojot korpusu un citu valodas resursu datus, valodnieks var precīzāk definēt vārda nozīmi un nozīmes nianses, ja redz to vairākos piemēros vienkopus, turklāt vienlaikus var izvēlēties raksturīgākos piemērus, ko rādīt vārdnīcas mikrostruktūrā. Salīdzinājumā ar agrākiem gadiem, kad piemēri tika meklēti drukātos izdevumos un kataloģizēti kartotēkās, šāds datu ieguves ceļš daudzkārt atvieglo valodnieku pētījumus un arī lietotāju meklējumus. Taču patlaban, kad datu apjoms arvien pieaug, jāatceras, ka pārāk daudz datu var apgrūtināt tā analizētāju, tāpēc digitālajām humanitārajām zinātnēm nepārtraukti jārūpējas par datu apstrādi un to digitālo ekspozīciju.

Vārdnīcu digitalizēšana gandrīz nepārtraukti paplašina lietotāja iespējas un ērtu piekļuvi – lasītājam vairs nav jāšķirsta drukāti izdevumi un jāpārzina alfabēts, mijnorāžu atrašanai vien jāatver saite, kontekstu var iepazīt ne tikai sintagmu, teikumu vai rindkopu, bet pat diskursa apjomā. Elektroniskas vārdnīcas lietotājam vairs nav jāmeklē saīsinājumu saraksts, jo digitālo lauku telpa nav tik ļoti ierobežota kā drukātam izdevumam. Vārdnīcas vairāk nekā jebkad agrāk tiek pielāgotas lietotāja vajadzībām, un šo vajadzību apmierināšana tiek uzskatīta par nākotnes vārdnīcu mērķi. Lietotāju vēlmes konstatētas jau elektroniskās leksikogrāfijas aizsākumos: ir noskaidrots, ka vārdnīcas lasītājs galvenokārt

interesējas, vai šāds vārds vispār eksistē, kā to raksta un izrunā un kāda ir tā nozīme vai nozīmes nianses (Wallraff 2009). Par elektroniskas vārdnīcas pielāgošanu jādomā jau tās izveides sākumposmā, kad datu ievadīšanas posms vēl nav sācies. Arvien vairāk elektronisko leksikogrāfu izvēlas radīt pēc iespējas vairāk valodas digitālās ekspozīcijas formu, lai vēlāk vārdnīcu varētu pielāgot vai pārveidot: tiek anotētas un kodētas valodas vienību daļas, sākot no plašākiem tekstiem līdz pat fonēmām; tiek pievienota izruna audioierakstā vai starptautiskā fonētiskā sistēmā; tiek veidotas pilnas vārda gramatiskās paradigmas utt. Šāda elektronisko vārdnīcu arhitektūras izstrāde tālāk var rosināt valodas resursu hibridizāciju, kurā, kā plāno digitālo humanitāro zinātņu pētnieki, robežas starp vārdnīcu, rīku un korpusu palēnām vājināsies vai pat izzudīs (Granger, Paquot 2012: 5) un tie visi taps par vienotu valodas resursu, proti, vienā resursā varēs atrast visu par konkrētu valodas vienību. Latvijas elektroniskajā leksikogrāfijā dažādu valodas resursu robežas vēl skaidri iezīmējas, taču, piemēram, LVVV ciešā sasaiste ar korpusu un vietnes *tezaurs.lv* datu akumulēšana no visām elektroniskajām vārdnīcām jau liecina par daļēju hibridizāciju arī latviešu humanitārajās zinātnēs.

Hibridizāciju arvien var uzlabot digitālo resursu sistēmiskums. Vienota sistēma un sistēmiskums ir visu digitālo humanitāro zinātņu pamatā, un tas ir galvenais digitālo humanitāro zinātņu devums humanitārajām zinātnēm, arī elektroniskajai leksikogrāfijai. Digitālā vide prasa proformas, ko pielāgot lieliem valodas resursiem, un īpaši elektroniskajā leksikogrāfijā tāas tiek izstrādāti un izmantotasikviena šķirkļa izveidē. Līdz ar to pārdomāti izstrādāta digitālā vide atšķirībā no drukāta izdevuma nekad neļaus pazaudēt kādu no sistēmas pamatelementiem, neļaus aizmirst nevienu obligāto informācijas vienības aspektu (piemēram, TVD vietvārdu nevar ievadīt, ja nav norādīta precīza tā atrašanās vieta administratīvā teritorijā, AD apvidvārdu nevar pievienot, ja nav norādīta tā nozīme un/vai lokalizācija).

Šādas obligāto elementu sistēmas ir ļoti svarīgas pūļpakalpes rīkos, kur leksikogrāfiskos laukus var aizpildīt jebkurš lietotājs. Latvijā lietotāju līdzdalība leksikogrāfisku avotu izstrādē tiek īpaši veicināta. Digitālais laikmets nevar pastāvēt bez nemitīgas sasaistes starp lietotāju un izstrādātāju. Elektronisku vārdnīcu izstrādātājiem nemitīgi jāinteresējas par lietotāju vēlmēm un palaikam jāatļauj lingvistisko leksikogrāfiju veidot arī pašiem. Sabiedrība tiek iesaistīta ne tikai elektronisko vārdnīcu šķirkļu rediģēšanā (piemēram, MLVV un LLVV ir iespējams ziņot par kādu neprecizitāti valodnieku izstrādāto šķirkļu elementos), bet arī vārdnīcas vai datubāzes satura veidošanā. LU Latviešu valodas institūtā ir izstrādāti divi pūļpakalpes rīki – leksikogrāfiski avoti, kas balstās tikai uz lietotāju iesūtītās informācijas datiem (“Tautas vietvārdu datubāze” un “Apvidvārdu datubāze”).

Šādi elektroniski leksikogrāfiski rīki sniedz ne tikai citādi grūti iegūstamus valodas datus, bet arī rāda, ka sabiedrība aptver savas valodas faktu vērtību un apzinās, ka tie jāsargā un jādara zināmi citiem.

Digitalizācija nodrošina ne tikai datu pieejamību, pārskatāmību, bet arī dabas saudzēšanu, ko uzsver ikviens digitālā laikmeta pētnieks.

Secinājumi

LU LaVI pētnieku darbs veido trīs lielāko latviešu valodas elektronisko vārdnīcu pamatu: LLVV, MLVV, LVVV. Tās aptver latviešu valodas pamatkorpusu gan diahroniski, gan sinhroniski: LVVV tiecas kļūt par seno rakstu avotu tēzauru, LLVV aptver literārās valodas vārdu kopumu, savukārt MLVV pamazām tiek publicēti visi mūsdienās lietotie un arvien no jauna radušies vārdi un vārdu savienojumi. Pēdējos gados LU LaVI izstrādātie elektroniskās leksikogrāfijas avoti tiecas aptvert arī Latvijas reģionos lietoto valodu: AD un TVD mērķis ir uzkrāt literārajā valodā neiekļautos, izloksnēs lietotos vārdus.

Iepazīstot digitālo humanitāro zinātņu pētnieku atziņas, pieredzi un paraugus, secināms, ka latviešu elektroniskajā leksikogrāfijā vēl nepieciešami uzlabojumi. Pirmkārt, latviešu elektronisko vārdnīcu šķirkļu izstrādē trūkst proformu (šķirkļu veidnes), kas atvieglotu gan vārdnīcu sastādītāju darbu, atgādinot par obligātām formulējumu struktūras un semantikas sadaļām, gan ļautu lietotājam vieglāk uztvert vārdu nozīmes un gūt pilnvērtīgāku vārda nozīmes skaidrojumu. Otrkārt, lai gan meklēšanas iespējas elektroniskajās vārdnīcās ir gandrīz neierobežotas, visām vārdnīcām būtu nepieciešams leksēmu alfabētisks rādītājs, kas lietotājam ļautu atrast vārdu, kura formu viņš precīzi nezina, turklāt radītu pilnīgāku priekšstatu par valodas sistēmu. Treškārt, lai atspoguļotu reālo valodu un valodas būtību, elektroniskajām vārdnīcām būtu nepieciešams pievienot audiomateriālus (vārda reālo izrunu) – apjaušot to nozīmību, iespēja ierunāt vārdu ir iekļauta "Apvidvārdu datubāzē". Ceturtkārt, nepieciešams arvien paplašināt valodu korpusus un tos integrēt elektronisko vārdnīcu izstrādē, lai nodrošinātu precīzāku šķirkļu izstrādi.

Abel, Andrea (2012). Dictionary writing systems and beyond. Granger, Sylviane; Paquot, Magali (eds.). *Electronic Lexicography.* Oxford: Oxford University Press, pp. 83–106.

Abney, Steven (2011). *Language Digitization*. Write-up of a talk given at the University of Michigan. Available: http://www.vinartus.net/spa/p102-v2.pdf [accessed 01.11.2011.].

AD – *Apvidvārdu datubāze.* Pieejams: https://apvidvardi.lu.lv [skatīts 17.05.2022.].

Atkins, Beryl T. Sue, Rundell, Michael (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Beal, John, Corrigan, Karen, Moisl, Hermann (eds.) (2007). *Creating and digitizing language corpora*, Vol. 1. *Synchronic databases*. Hampshire&New York: Palgrave Macmillan.

Bird, Steven, Simons, Gary (2003). Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. *Computers and the Humanities*, 37 (4), pp. 375–388.

Burdick, Anne, Drucker, Johanna, Lunenfeld, Peter, Presner, Todd, Schnapp, Jeffrey. *Digital Humanities.* Cambridge, London: MIT Press, 2012.

Fuertes-Olivera, Pedro A., Bergenholtz, Henning (2013). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Continuum.

Jērāne, Santa (2014). Vietvārdu vārdnīcu megastruktūra. *Linguistica Lettica*, Nr. 22. Rīga: LU Latviešu valodas institūts, 185.–204. lpp.

Jērāne, Santa (2021). Cirkulāri skaidrojumi "Mūsdienu latviešu valodas vārdnīcā": dažas teorētiskas un praktiskas problēmas. *Linguistica Lettica*, Nr. 28. Rīga: LU Latviešu valodas institūts, 164.–173. lpp.

Lew, Robert (2010). Multimodal Lexicography: The Representation of Meaning in Electronic Dictionaries. *Lexikos*, No. 20, pp. 290–306.

Logan, Harry M. (1991). Electronic Lexicography. *Computers and the Humanities*, No. 25(6), pp. 351–361.

LLVV = *Latviešu literārās valodas vārdnīca* [elektroniska vārdnīca]. Pieejams: https://llvv.tezaurs.lv/ [skatīts 17.05.2022.].

LVVV = *Latviešu valodas vēsturiskā vārdnīca* [elektroniska vārdnīca]. Pieejams: https://tezaurs.lv/lvvv/ [skatīts 17.05.2022.].

MLVV = *Mūsdienu latviešu valodas vārdnīca* [elektroniska vārdnīca]. Pieejams: https://mlvv.tezaurs.lv/ [skatīts 17.05.2022.].

Roze, Liene (1982). *Pasaule vārdnīcas skatījumā.* Rīga: Zinātne.

Granger, Sylviane; Paquot, Magali (2012). *Electronic Lexicography.* Oxford: Oxford University Press.

Spohr, Dennis (2011). A multi-layer architecture for 'pluri-monofunctional' dictionaries. *e-Lexicography – The Internet, Digital Initiatives and Lexicography*. London and New York: Continuum, pp. 103–120.

TVD – *Tautas vietvārdu datubāze*. Pieejams: www.vietvardi.lv [skatīts 17.05.2022.].

Wallraff, Barbara (2009). The uncertain future of dictionaries. *The Atlantic*, 12 January. Available: http://barbarawallraff.theatlantic.com/archives/2009/01/the_uncertain_future_of_dictio.php [accessed 12.12.2021.].

Won, Miguel. Murrieta-Flores, Patricia, Martins, Bruno (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, Vol. 5:2. Available: https://www.frontiersin.org/articles/410.3389/fdigh.2018.00002/full [accessed 30.10.2021.].

Weschler, Robert, Pitts, Chris (2000). An experiment using electronic dictionaries with EFL students. *The Internet TESL Journal*, No. 6. Available: www.iteslj.org/ Articles/Weschler-ElectroDict.html [accessed 28.10.2021.].

Summary

The Digital Structure of Dictionary

Sanda Rapa

Keywords: e-lexicography, electronic dictionary, microstructure of e-dictionary, macrostructure of e-dictionary, history of Latvian e-lexicography

E-lexicography has not got long history in Latvia: the first electronic dictionaries appeared in the late 1990s. However, many achievements can be traced over the past twenty years: electronic dictionaries of Early Latvian (https://tezaurs.lv/lvvv), Standard Latvian (https://llvv.tezaurs.lv/) and Modern Latvian (https://mlvv.tezaurs.lv/) have been published. Also, databases for crowdsourcing of Latvian place names (www.vietvardi.lv) and regional words (https://apvidvardi.lu.lv) have been created. These resources comprise more than 100 thousand words each of which is accompanied by a linguistic description, examples and data of registration. All of the e-lexicografic sources in Latvian are elaborated in cooperation between the Latvian Language Institute and the Institute of Mathematics and Computer Sciences of the University of Latvia. Based on this experience, the article deals with the history, specificity and importance of the digitization of Latvian language from the point of view of content creators of digital tools.

The article is structured into three subchapters. The first part provides an insight into the history of Latvian e-lexicography, the second part analyses the principles of compiling the electronic lexicographic sources in Latvia (including the analysis of microstructure and macrostructure of published electronic dictionaries). The third part deals with the pros and cons of Latvian lexicographic sources as opposed to printed sources (the analysis of corpus integration, data sufficiency and unification, efficiency of access, customization, hybridization, user input is provided). The aim of the article is to outline priorities for the future of Latvian electronic lexicography.

The analysis of Latvian electronic dictionaries reveals four areas where Latvian e-lexicography still requires an improvement. Firstly, there is a lack of proformas in all electronic dictionaries of Latvian language which could help to unify the structure of definition and meaning of the entries. Secondly, not all Latvian electronic dictionaries comprise an alphabetical list of lemmas which could help to find an entry without knowing the spelling of the word. Thirdly, audio pronunciation for each entry is necessary in order to provide real (standardized) pronunciation. Fourthly, corpus integration is essential to produce rich lexical entries and to define the meaning of lemmas more precisely.