Gunta Nešpore, Normunds Grūzītis,
Everita Andronova, Andrejs Spektors

K. MĪLENBAHA UN J. ENDZELĪNA "LATVIEŠU VALODAS VĀRDNĪCAS" PILNVEIDOTA ELEKTRONISKĀ VERSIJA

K. Mīlenbaha un J. Endzelīna "Latviešu valodas vārdnīca" (ME) ir unikāls latviešu valodas faktu materiāls, kas paver plašas iespējas valodas pētīšanā. Latviešu valodā nav otra tik plaša apvidvārdu un vecvārdu krājuma. Arī pēc formas "Latviešu valodas vārdnīca" ir visai neparasta – tā reizē ir gan skaidrojoša, gan tulkojoša vārdnīca, tai raksturīgas arī etimoloģijas un sinonīmu vārdnīcas iezīmes. Līdz 1992. gadam, kad iznāca K. Karuļa "Latviešu etimoloģijas vārdnīca", ME bija galvenais avots, kurā atrodamas ziņas par latviešu valodas vārdu cilmi; arī pēc tam tās nozīme nav mazinājusies.

Lai padarītu šo vārdnīcu plašāk pieejamu (ME var uzskatīt par bibliogrāfisku retumu), ērtāk un daudzveidīgāk lietojamu (nesalīdzināmi plašākas meklēšanas iespējas), Latvijas Universitātes Matemātikas un informātikas institūtā ir izveidota K. Mīlenbaha un J. Endzelīna "Latviešu valodas vārdnīcas" elektroniskā versija.

ME ELEKTRONISKĀS VERSIJAS IZVEIDES AIZSĀKUMS

ME ievadīšana datorā aizsākās 1994. gadā, ievadot vārdnīcas tekstu datorā ar rokām. Šādā veidā darbs ritēja lēni, tam nebija arī atsevišķa finansējuma. Situācija būtiski uzlabojās 2000. gadā, kad ar Latviešu fonda (ASV) finansiālu atbalstu tika elektronizēti visi vārdnīcas pamatsējumi, kā arī izstrādāti meklēšanas sistēmas pamati. Šajā laikā tīmeklī tika ievietota elektroniskās vārdnīcas izmēģinājuma versija (A, Ā burts).

Viss vārdnīcas pamatsējumu teksts pirmajā versijā ar meklēšanas iespējām tīmeklī bija pieejams 2002. gadā.

Tā kā ME ir izmantotas ļoti daudzas specifiskas rakstzīmes, sākotnēji tās tika ievadītas datorā, izmantojot īpaši izveidotus apzīmējumus (piemērus skatīt 1. attēlā). Šie apzīmējumi bija ērti darba gaitā, bet ne tik ērti un viegli uztverami vārdnīcas elektroniskās versijas lietotājiem, tāpēc laika gaitā radās nepieciešamība meklēt tehniskus risinājumus, lai varētu pilnīgāk attēlot vārdnīcas tekstu.

Tehnisku iemeslu dēļ sākotnēji netika attēloti teksta fragmenti, kas nav latīņu alfabētā, tie tika aizstāti ar zīmi #.

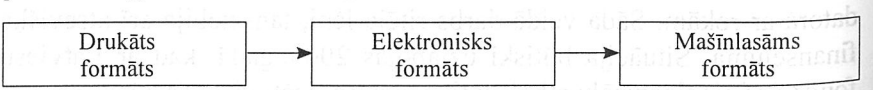
a^ – lauztā zilbes intonācija (**â, ê, ...**)
a~ – cirkumflekss (**ã, ẽ, ...**)
a” – umlauts (**ë, ä ...**)
a. – palīgakcents (**a˙**)
l/ – poļu valodas **ł**
a< – akūts (**á, é, ...**)
a& – lietuviešu un poļu valodas nazālie patskaņi (**ą, ę, ...**)
e| – lietuviešu valodas **ė**

1. attēls. Vārdnīcas elektronizēšanā izmantoto apzīmējumu piemēri

Lai paplašinātu meklēšanas iespējas vārdnīcas elektroniskajā versijā, katram šķirklim tika pievienots šķirkļa vārds mūsdienu ortogrāfijā, piemēram, šķirklim, kurā aprakstīts vārds *ķŗẽguôt*, tika pievienots vārds *krēgot*.

Marķējums

Lai izceltu teksta (šajā gadījumā – vārdnīcas) formāla strukturālā marķējuma pievienoto vērtību, aplūkosim vienkāršotu formātu attīstības posmu diagrammu:



Mašīnlasāms formāts ir būtisks solis pievienotās vērtības celšanā, jo šādi atšķirībā no vienkārši elektroniska formāta (*.txt*, *.rtf*, *.doc* u. c.)

ar formāliem līdzekļiem tiek precīzi norādīta teksta vienības (šķirkļa) iekšējās struktūras uzbūve – loģiskie elementi, to secība un savstarpējais izkārtojums, organizācija. Tas savukārt paver plašas iespējas vārdnīcas datorizētā apstrādē.

Projekta pirmsākumos, 1994. gadā, ME šķirkļu marķēšanai tika ieviesta lielākoties reprezentatīvu iezīmju kopa (ap 10 dažādu elementu aprakstīšanai) – procedurāls marķējums – šķirkļu struktūras elementi tika aprakstīti, galvenokārt ņemot vērā vārdnīcas oriģināla noformējumu. Vairāku būtisku iemeslu dēļ šāds marķējums (nedaudz papildināts) ir izmantots arī visos turpmākajos projekta attīstības posmos.

- Viens no sākotnējiem marķējuma mērķiem – maksimāla atbilstība vārdnīcas oriģinālam.
- Marķēšana tika veikta gan manuāli, gan daļēji automatizēti (analizējot RTF[1] formātā ieskenētās burtnīcas); līdz 2000. gada posmam šādi bija marķēti vārdnīcas pirmie pamatsējumi.
- Kopumā šķirkļu iekšējā struktūra ir ļoti komplicēta – automātiski atpazīt šķirkļa elementus ir problemātiski.
- Izmantoti atsevišķi daudznozīmīgi strukturālie elementi; virspusējā šķirkļu struktūra tika atpazīta pēc procedurālajām iezīmēm.

@m pussvētvakars	// Šķirkļa vārds mūsdienu ortogrāfijā
@1 pussve`,tvakars,	// Šķirkļa vārds ME pierakstā
@v der Sonnabendabend	// Teksts vācu valodā
@2 Dond. n. RKr. XVII, 48:	// Teksts sīkā drukā
@4 sestdienas vakars ir pussvē,tvakars.	// Teksts kursīvā
@z ME III, 435	// Avota norāde

2. attēls. Marķēta šķirkļa piemērs

Papildus katra šķirkļa beigās ir norādīts sējums un lappuse, kurā tas atrodams vārdnīcā; vārdnīcas beigās pievienotie kļūdu labojumi ir piesaistīti attiecīgajiem šķirkļiem. Marķējuma piemērs ir aplūkojams 2. attēlā.

Līdz ar to ME formāts tikai nosacīti atbilst iepriekš apskatītajām mašīnlasāma formāta prasībām, tomēr arī šādas ierobežoti lietotas iezīmes jau sniedz iespēju ar lielu varbūtību atpazīt dažādus šķirkļa elementus.

[1] RTF – *Rich Text Format*

ME ELEKTRONISKĀS VERSIJAS PILNVEIDOŠANA

Pirmā ME elektroniskā versija padarīja to pieejamu tīmeklī un piedāvāja lietotājiem plašas meklēšanas iespējas, tomēr tai bija vairākas nepilnības, kuru novēršana padarītu šo vārdnīcu ērtāk lietojamu, kā arī atspoguļotu elektroniskajā versijā trūkstošo ME informāciju.

2004. gadā, balstoties uz esošajām iestrādēm, ar Izglītības un zinātnes ministrijas finansiālu atbalstu kopējai sistēmai tika pievienoti "Latviešu valodas vārdnīcas" papildsējumi (EH). ME un EH šķirkļi meklēšanas sistēmā tiek atspoguļoti vienlaidus, taču tos ir viegli atšķirt gan pēc EH šķirklī izmantotā apzīmējuma, gan pēc avota norādes šķirkļa beigās; turklāt EH šķirkļi sarakstā vienmēr ir doti pēc attiecīgajiem ME šķirkļiem.

Šajā posmā nozīmīgs darbs tika veikts arī vārdnīcas kvalitātes uzlabošanā, izstrādājot programmatūru, ar kuras palīdzību daļēji automātiski tika novērsts ievērojams apjoms ievadkļūdu un marķējuma kļūdu:

- tika izmantotas regulāras izteiksmes, lai atpazītu korektus šķirkļa elementus un saņemtu brīdinājumu par "aizdomīgiem" elementiem;
- tika pārbaudīta šķirkļa elementu (iezīmju) secība u. tml.

Jāpiemin arī kļūdu novēršanas procesa papildu ieguvums: "atradās" ap 2700 pazudušu šķirkļu, kas līdz šim nebija atrodami dažādu strukturālā marķējuma kļūdu dēļ vai arī bija "saplūduši" ar citiem šķirkļiem.

Bez papildsējumu pievienošanas vārdnīcas pilnveidotās elektroniskās versijas būtiskākais jauninājums ir pāreja no vārdnīcā līdz šim izmantotā ierobežotā baltu valodu rakstzīmju kodēšanas standarta[2] uz *Unicode*[3] – vienotu standartu, kas aptver praktiski visas pasaules valodas: gan mūsdienu, gan vēsturiskās.

Šāds solis pavēra iespēju korekti attēlot ne tikai izmantotās specifiskās rakstzīmes teksta fragmentos latviešu un lietuviešu valodā, bet arī ievadīt trūkstošos citu valodu (krievu, grieķu valodas, sanskrita u. c.) piemērus.

---

2 *Baltic (Windows) code page (Cp1257)*

3 http://www.unicode.org/charts

Vispirms tika sastādīts pilns ME un EH izmantotais "alfabēts" – 218 rakstzīmju saraksts. Šī saraksta elementiem tika piekārtoti atbilstošie apzīmējumi (simbolu kombinācijas), kas tika izmantoti līdzšinējā vārdnīcas kodēšanā, iegūstot substitūcijas tabulu (skatīt 3. attēlu). Tā kā specifisko rakstzīmju kodēšana un atainošana ir tehniski sarežģīta, liela daļa svešvalodu fragmentu tika ievadīti ar dažādu apzīmējumu palīdzību.

Nākamais solis bija *Unicode* izpēte, mēģinot šī standarta tabulās atrast kodus visām 218 rakstzīmēm. Izrādījās, ka vārdnīcā izmantotie apzīmējumi ir tik bagātīgi, ka 46 % rakstzīmju nācās kodēt ar 2–3 simbolu kombinācijām tā, lai tiktu iegūti simboli, kas apzīmē vienu skaņu.

Visbeidzot tika izveidota konvertēšanas programmatūra, kas, izmantojot minēto substitūcijas tabulu, pārveidoja vārdnīcu korektā *Unicode* formātā un aizstāja esošos tiešsaistes datubāzes šķirkļus ar jaunajiem.

Rakstzīmju standartizēta kodēšana vēl neatrisina visas ar vārdnīcas rakstzīmēm saistītās problēmas – paliek vēl to atainošanas, tātad fontu problēmas. Pirmkārt, lai atainotu simbolus, ir jāizvēlas tāds fonts, kas būtu pieejams ikvienam potenciālajam vārdnīcas elektroniskās versijas lietotājam. Otrkārt, šādam fontam ir jāatbalsta *Unicode* standarts un jāpārklāj ME un EH izmantoto rakstzīmju kopa. Ikdienā visbiežāk izmantotie un plaši pieejamie fonti, kā *Times New Roman* vai *Arial,* tikai daļēji atspoguļo nepieciešamās *Unicode* apakškopas un nepilnīgi atbalsta rakstzīmju kombinēšanu.

Veicot izpēti plaši vai bez maksas pieejamo *Unicode* fontu klāstā, šobrīd ir izraudzīti divi vārdnīcas atainošanai izmantojamie fonti (pēc lietotāja izvēles): *Arial Unicode MS* un *Doulos SIL*. Pirmais ir pieejams visiem *MS Office 2000* (vai jaunāka) lietotājiem un aptver plašu *Unicode* daļu, līdz ar to arī tā izmērs ir diezgan liels (~23 MB). Savukārt otrs ir kompromisa variants: tīmeklī lejupielādējams bez maksas, salīdzinoši neliels (~800 KB), atbalsta rakstzīmju kombinēšanu, taču nepilnīgi aptver atsevišķu valodu (piemēram, grieķu) alfabētus. Abu fontu gadījumā bija jāatsakās no atsevišķām (par laimi reti lietotām) rakstzīmju kombinācijām, piemēram, *l* ar stieptās intonācijas apzīmējumu, jo fontu augstuma ierobežojumi neļauj korekti atainot tildi virs *l*; šādos gadījumos ir saglabātas divu pozīciju kombinācijas (skatīt 3. attēlu). Tomēr pamatā

daudzveidīgās vārdnīcā izmantotās rakstzīmes pašlaik ir atveidotas korekti, un lietotājs tās var ērti lasīt un uztvert bez grūtībām.

Šobrīd sākumā ieviestie rakstzīmju apzīmējumi var tikt izmantoti, ievadot meklēšanas pieprasījumu, meklējot pēc teksta fragmenta ME oriģinālrakstībā.

Pašreizējā situācija:

- tiešsaistē ir pieejama gan ME, gan EH;
- ir adekvāti attēlota praktiski visa vārdnīcā esošā informācija;
- pamatsējumos atrodami **77 175** šķirkļi, papildsējumos — **58 238** šķirkļi;
- tīmekļa vietne: **www.ailab.lv/MEV** (šeit ir atrodams meklēšanas iespēju apraksts, informācija par tehniskajām prasībām u. tml.; pagaidām elektronisko vārdnīcu var izmantot tikai ar *Microsoft Internet Explorer* pārlūkprogrammu).

No			Uz		
Reģistrs	**Zīme**	**Unicode**	**Reģistrs**	**Zīme**	**Unicode**
lower	ē,<	ē+,+<	lower	ḗ	1E17+0326
lower	l,~	l+,+~	lower	ɭ̃	l+0327+02DC
lower	ņ^	ņ+^	lower	ņ̂	ņ+0302
lower	y~	y+~	lower	ỹ	1EF9
lower	ā¯	00E4+02C9	lower	a̅̈	01DF
...	...	...	...	...	...

3. attēls. Rakstzīmju apzīmējumu substitūcijas tabula

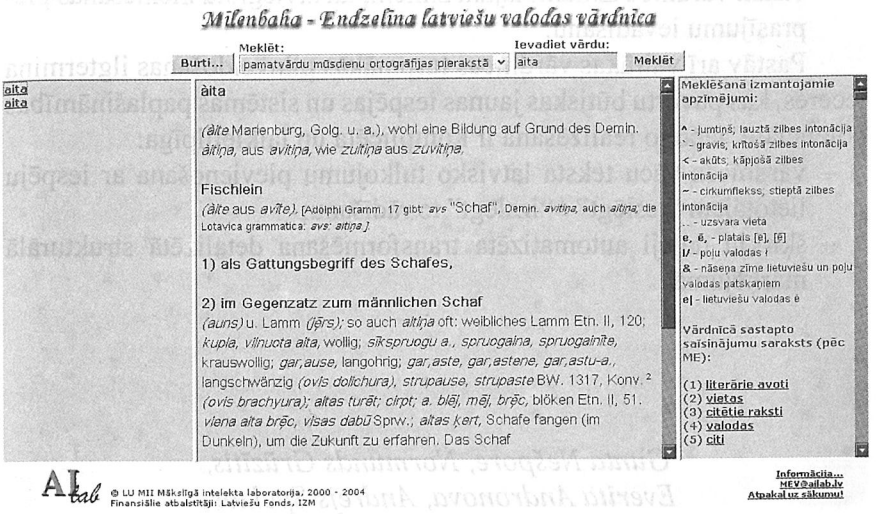
MAZLIET PAR MEKLĒŠANAS IESPĒJĀM VĀRDNĪCAS ELEKTRONISKAJĀ VERSIJĀ

Viena no būtiskākajām vārdnīcas elektroniskās versijas priekšrocībām, salīdzinot ar "papīra" vārdnīcu, ir plašas meklēšanas iespējas. Vārdnīcas lietotāja saskarne (skatīt 4. attēlu) ir izveidota tā, lai lietotājs pēc iespējas ērtāk varētu atrast un izmantot visu viņam nepieciešamo informāciju. Vienlaikus redzami gan atrastie šķirkļi, gan meklēšanas

pieprasījumā izmantojamo apzīmējumu saraksts, gan vārdnīcā sastopamie saīsinājumi un cita informācija.

Pēc lietotāja izvēles meklēt vārdnīcā var:

- šķirkļa vārdu vai tā daļu (mūsdienu ortogrāfijā vai ME izmantotajā pierakstā);
- teksta fragmentu šķirkļu skaidrojumos;
- teksta fragmentu kļūdu labojumos;
- šķirkļus, kas atrodami noteiktā vārdnīcas sējumā un lappusē (pēc avota).



4. attēls. Vārdnīcas lietotāja saskarne

Daudzveidīga informācija ir atrodama, meklējot teksta fragmentus ar šablonu palīdzību. Ar zīmi % tiek apzīmēts jebkāda garuma teksta (vārda) fragments, bet ar zīmi _ — jebkurš viens simbols. Piemēram, meklēšanas pieprasījuma logā ievadot *ap%ties*, tiek atrasti visi vārdi, kas sākas ar *ap* un beidzas ar *ties*; ievadot pieprasījumu *a_i__a%*, tiek atrasti vārdi, kas sākas ar burtu *a*, kura otrais burts var būt jebkurš, trešais ir *i*, ceturtais un piektais var būt jebkurš, sestajam jābūt *a*, tālāk var sekot jebkas.

Papildus lietotājs var norādīt, kādos burtos vārdnīcā veikt meklēšanu (poga *Burti*). Tad sistēma meklēs tikai tajos šķirkļos, kuru pamatvārds sākas ar kādu no norādītajiem burtiem.

NĀKOTNES IECERES

Tuvākajā nākotnē ir plānoti vairāki funkcionāli elektroniskās vārdnīcas jauninājumi:

- šķirkļu sasaiste ar vārdnīcas oriģināla faksimiliem, lai vajadzības gadījumā tīmekļa versijas lietotājiem būtu iespēja salīdzināt interesējošo šķirkli ar oriģinālu; vārdnīcas faksimili šobrīd jau ir sagatavoti, tikai vēl nav pievienoti elektroniskajai vārdnīcai;
- ekrāna tastatūras pievienošana vārdnīcas lietotāja saskarnei ar visām vārdnīcā izmantotajām zīmēm, lai atvieglotu meklēšanas pieprasījumu ievadīšanu.

Pastāv arī vairākas vārdnīcas turpmākas pilnveidošanas ilgtermiņa ieceres, kas pavērtu būtiskas jaunas iespējas un sistēmas paplašināmības potenciālu, taču to realizēšana ir komplicēta un laikietilpīga:

- vārdnīcas vācu teksta latvisko tulkojumu pievienošana ar iespēju lietotājam "ieslēgt" / "izslēgt" to rādīšanu;
- šķirkļu daļēji automatizēta transformēšana detalizētā strukturālā marķējumā.

Gunta Nešpore, Normunds Grūzītis,
Everita Andronova, Andrejs Spektors

IMPROVED ELECTRONIC VERSION OF K. MÜLENBACHS' AND J. ENDZELĪNS' "LETTISCH-DEUTSCHES WÖRTERBUCH"

Summary

K. Mülenbachs' and J. Endzelīns' "Lettisch-deutsches Wörterbuch" (ME) contains a unique treasure of the Latvian language, offering wide scope opportunities in the language study. It covers a large amount of dialect words and lexis of Old Latvian. "Lettisch-deutsches Wörterbuch" has a quite unusual composition – it is both an explanatory dictionary and a Latvian-German

dictionary, reflecting the peculiarities of an etymological and synonym dictionary as well. Up to 1992, when "The Latvian Etymological Dictionary" compiled by K. Karulis was published, ME was the main source to find the information about the word origin. The ME dictionary has not lost its topicality till now.

In order to make this dictionary more open to the public (ME has already become a bibliographical rarity) and to ensure more convenient use (with wide search possibilities), the Institute of Mathematics and Computer Science, University of Latvia has improved the previously developed electronic version of the dictionary (visit www.ailab.lv/mev).