https://doi.org/10.22364/vnf.11.01

Mērķhipotēžu izvirzīšana latviešu valodas apguvēju korpusā

Creating target hypotheses in a learner corpus of Latvian

Ilze Auziņa, Kristīne Levāne-Petrova
Mākslīgā intelekta laboratorija
Matemātikas un informātikas institūts, Latvijas Universitāte
Raiņa bulvāris 29, Rīga, LV-1459, Latvija
E-pasts: *ilze.auzina@lumii.lv, kristine.levane-petrova@lumii.lv*

Inga Kaija
Rīgas Stradiņa universitāte
Dzirciema iela 16, Rīga, LV-1007

Mākslīgā intelekta laboratorija
Matemātikas un informātikas institūts, Latvijas Universitāte
Raiņa bulvāris 29, Rīga, LV-1459, Latvija
E-pasts: *inga.kaija@rsu.lv*

Apguvēju korpuss ir sistemātiski datorizētu valodas apguvēju (gan svešvalodas, gan otrās valodas) veidotu tekstu datubāze. Tas ir ārvalstnieku valodas apguvēju īpatnību izpētes un datos balstītu latviešu valodas mācību materiālu un metodisko līdzekļu izstrādes pamats. Apguvēju korpusu, tāpat kā citus valodas korpusus, var marķēt dažādos valodas līmeņos (morfoloģiski, sintaktiski), bet īpaši nozīmīgs apguvēju valodas izpētē ir kļūdu marķējums un tajā balstītā kļūdu analīze. Kļūdu analīzi ietekmē divi faktori: 1) izraudzītie kļūdu tipi jeb kļūdu tipoloģija un 2) izvirzītās mērķhipotēzes, t. i., labotais teksts. Tādēļ pirms kļūdu marķēšanas ir būtiski vienoties, kas tiks marķēts un kā tas tiks darīts.
Raksta ievadā ir īsi raksturots veidojamais „Latviešu valodas apguvēju korpuss" (LaVA), aplūkots mērķhipotēzes jēdziens un mērķhipotēzes nozīme valodas apguvēju korpusa izveides procesā. Rakstā ir izklāstīti galvenie mērķhipotēzes izvirzīšanas principi korpusā LaVA, kā arī minēti konkrēti piemēri, kā valodas apguvēju izteikumi tiek laboti atbilstoši latviešu valodas normām un kādas ir būtiskākās atkāpes, kas tiek pieļautas.

Atslēgvārdi: valodas korpuss; valodas apguvēju korpuss; mērķhipotēze; kļūdu marķēšana; valodas apguve; korpuslingvistika.

Ievads

Latviešu valodas kā svešvalodas mācīšana kļūst aizvien populārāka Latvijas un arī ārvalstu augstākajās mācību iestādēs, tāpēc jaunu, korpusā balstītu mācību materiālu izstrāde ir ļoti nozīmīga (Laizāne 2019). Lai pētītu ārvalstnieku latviešu valodas apguves īpatnības un nodrošinātu datos balstītu latviešu valodas mācību un metodisko materiālu izstrādi, ir nepieciešams latviešu valodas apguvēju korpuss.

Šim nolūkam tiek veidots „Latviešu valodas apguvēju korpuss" (turpmāk tekstā – LaVA)[1].

Korpuslingvistikā par valodas apguvēju korpusu sauc sistemātiski datorizētu valodas apguvēju (gan svešvalodas, gan otrās valodas) veidotu tekstu datubāzi (Leech 1998, xiv; Nesselhauf 2005, 40). Tie ir specializēti korpusi, kuros apkopoti valodas apguvēju radīti teksti. Pašlaik valodas apguvēju korpusu jomā dominē angļu valodas apguvēju korpusi, piem., „International Corpus of Learner English v2" (ICLE v2) (Granger et al. 2009), „Louvain International Database of Spoken English Interlanguage" (LINDSEI)) (Gilquin et al. 2010), „Viena-Oxford International Corpus of English" (VOICE), pieejams: https://www.univie.ac.at/voice/, tomēr arvien vairāk tiek veidoti arī citu valodas apguvēju korpusi, piem., vācu valodas apguvēju korpuss „Fehlerannotierte Lernerkorpus des Deutschen als Fremdsprache" (FALKO), pieejams: https://korpling.german.hu-berlin.de/falko-suche/ (Siemen et al. 2006), portugāļu valodas apguvēju korpuss „Learner Corpus of Portuguese L2" (COPLE2), pieejams: http://teitok.clul.ul.pt/cople2/index.php?action=home (Mendes et al. 2016), krievu valodas apguvēju korpuss „Russian Learner Corpus" (RLC), pieejams: http://web-corpora.net/RLC (Rakhilina et al. 2016).

LaVA iekļauti tādu Latvijas augstākās izglītības iestādēs studējošo darbi, kas latviešu valodu apgūst kā svešvalodu bez priekšzināšanām pirmo vai otro semestri (atkarībā no kursa satura aptuveni atbilst A1 vai A2 līmenim). Tie ir patstāvīgi rakstīti teksti par docētāja izvēlētu tematu. Apguvējiem rakstīšanas laikā ir atļauts izmantot dažādus palīglīdzekļus, izņemot mašīntulkotāju. Tekstu garumam (vārdu skaitam) katrā atsevišķā gadījumā nosacījumus izvirza docētāji, balstoties uz attiecīgā brīža pedagoģiskajām vajadzībām, taču ieteikums ir sasniegt vismaz 100 vārdu apjomu vienā tekstā. Daļa docētāju izvirza augstākas prasības, tāpēc daļa tekstu sasniedz 200 vārdus un pat pārsniedz tos. Kopumā korpusā paredzēts iekļaut vismaz 1000 šādu tekstu, tātad korpusa kopējais apjoms pārsniegs 100 000 vārdlietojumu. Katrs students iesniedz ne vairāk kā vienu tekstu semestrī, ne vairāk kā divus semestrus, tātad autoru skaits ir vairāk par 500 un mazāk par 1000. Līdz šim tajā ir iekļauti teksti, kurus rakstījuši Latvijas Universitātes, Rīgas Stradiņa universitātes, Liepājas Universitātes, Rēzeknes Tehnoloģiju akadēmijas un Latvijas Kultūras akadēmijas audzēkņi no dažādām valstīm. Tekstu autori pārstāv ļoti dažādas dzimtās valodas (vācu, somu, krievu, arābu u. c.), taču latviešu valodas kursos kā starpniekvaloda tiek izmantota angļu valoda.

Apguvēju korpuss, tāpat kā citi valodas korpusi, tiek marķēts gan morfoloģiski, gan sintaktiski, bet īpaši nozīmīgs apguvēju valodas izpētē ir kļūdu marķējums, kas parāda novirzes no valodas normas. Marķētās kļūdas palīdz apzināt problemātiskās jomas valodas apguves procesā. Kļūdu marķēšana ir sarežģīts uzdevums, īpaši jau tik morfoloģiski bagātai valodai kā latviešu valoda. Lai valodas apguvēju tekstos varētu marķēt kļūdas, teksti vispirms ir jāpārraksta atbilstoši valodas normām, t. i., jāizvirza tā saucamā mērķhipotēze. Lai mērķhipotēzes izvirzītu konsekventi,

[1] Korpuss tiek veidots Latvijas Zinātnes padomes Fundamentālo un lietišķo pētījumu projektā „Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums" (lzp-2018/1-0527) sadarbībā ar Valsts pētījumu programmas projekta „Latviešu valoda" Nr. VPP-IZM-2018/2-0002 apakšprojektu „Latviešu valodas apguve".

tādējādi atvieglojot tālāko kļūdu marķēšanas procesu, ir jāizstrādā vadlīnijas, kurās noteikts, kā valodas apguvēju teksti tiks laboti dažādos valodas līmeņos. Turpmāk rakstā tiks izklāstīti galvenie mērķhipotēzes izvirzīšanas principi korpusā LaVA, kā arī minēti konkrēti piemēri, kā valodas apguvēju izteikumi tiek laboti atbilstoši latviešu valodas normām un kādas ir būtiskākās atkāpes, kas tiek pieļautas. Rakstā mērķhipotēze un teksta labošana jeb pārrakstīšana atbilstoši normām tiek lietoti kā sinonīmi.

1. Mērķhipotēzes jēdziens

Kļūdu analīzi ietekmē divi faktori (James 1998, 106; Tono 2003, 804–805):

1) izraudzītie kļūdu tipi jeb kļūdu tipoloģija;
2) izvirzītās mērķhipotēzes, t. i., labotais teksts.

Tādēļ pirms kļūdu marķēšanas ir būtiski vienoties, kas tiks marķēts un kā tas tiks darīts, ņemot vērā vairākus aspektus (James 1998, 106–107; Tono 2003, 804; Znotiņa 2018, 94):

1) kļūdas tiek klasificētas atbilstoši valodas līmeņiem un gramatiskajām kategorijām (*linguistics category classification*), piem., gramatika (formveidošana) – darbības vārds – laika kategorija – personas kategorija;
2) kļūdas tiek klasificētas pēc izmaiņām attiecībā pret mērķvalodu (*target modification taxonomy*), konkrēti – izvirzīto mērķhipotēzi apgūstamajā valodā.

Atkarībā no tā, kādas izmaiņas valodas apguvēja izteikumā veiktas, tiek noteiktas trīs grupas (James 1998, 106–109; Tono 2003, 804):

1) izlaists valodas elements (angļu val. *omission*);
2) lieks valodas elements (angļu val. *addition*);
3) nepareiza vārdforma vai neatbilstoša leksēma (angļu val. *misformation*).

Karls Džeimss (*Carl James* 1998, 110–112) min vēl vairākas iespējamās grupas – kļūdaina secība (angļu val. *misordering*) un kombinētas kļūdas (angļu val. *blends*).

Kļūdu identificēšana un marķēšana katrā ziņā ir atkarīga no tā, vai valodas apguvēja izteikums ir atzīts par pareizu vai kļūdainu. Labotā forma jau kalpo kā norāde anotēšanas procesā: salīdzinot valodas apguvēja tekstu un laboto tekstu, var automātiski noteikt vismaz daļu kļūdu, piem., ortogrāfijas, interpunkcijas, formveidošanas kļūdas.

Lai valodas apguvēju korpusā marķētu kļūdas un nodrošinātu datorizētu kļūdu analīzi (angļu val. *computer-aided error analysis*, vairāk sk. Dagneaux et al. 1998, 163–174), viens no galvenajiem uzdevumiem ir interpretēt rakstīto, proti, izlemt, ko valodas apguvējs katrā konkrētā gadījumā ir gribējis teikt. Tāpēc korpuslingvistikā, runājot par valodas apguvēju korpusu izveidi, tiek piedāvāts mērķhipotēzes jēdziens (vācu val. *Zielhypothese*, angļu val. *target hypothesis*), kas ir „valodas apguvēju izteikumu rekonstrukciju mērķvalodā" (Ellis 1994, 54; sk. arī Lüdeling et al. 2005; Siemen et al. 2006, 131), ar kuru saprot priekšstatu, kā pareizi būtu veidota attiecīgā valodas struktūra, ņemot vērā ortogrāfijas, interpunkcijas, formveidošanas un vārddarināšanas principus. Mērķhipotēze nav absolūtā patiesība

vai vienīgais pareizais veids, kā kaut ko pateikt, bet tā ir izteikuma interpretācija konkrēta pētniecības mērķa sasniegšanai (Lüdeling et al. 2008, 68).

Par kļūdainām tiek uzskatītas tās teksta vienības, kuras neatbilst mērķhipotēzei, ko izvirza teksta labotājs, piem., teksta fragmenta *Es dzīvoju* ***Rīga*** mērķhipotēze ir *Es dzīvoju* ***Rīgā***, kurā ir vērojamā neatbilstība starp vārdformu *Rīga* (NOM SG) un *Rīgā* (LOC SG).

Skaidri formulēta mērķhipotēze ir nepieciešama, lai varētu veikt tālāku datu analīzi un marķēšanu. Lai arī vienam izteikumam ir iespējams neierobežots mērķhipotēžu skaits (Reznicek et al. 2013, 104–105) vai arī ir grūti vienoties par vienu variantu un ir korpusi, kuros tiek piedāvātas vairākas mērķhipotēzes, tomēr parasti valodas apguvēju korpusos katrai valodas vienībai tiek izvirzīta tikai viena mērķhipotēze. Arī LaVA korpusa uzbūve nedod iespēju izvirzīt vairākas mērķhipotēzes vienam un tam pašam teksta fragmentam, tāpēc par neviennozīmīgiem gadījumiem korpusa izstrādātāji vienojas, sniedzot iespējami ticamāko mērķhipotēzi.

LaVA korpusā katra teksta mērķhipotēze tiek izvirzīta divpakāpju procesā: vispirms divi korpusa veidotāji neatkarīgi viens no otra izveido katrs savu mērķhipotēzi, t. i., labo tekstu, un pēc tam trešais korpusa veidotājs īpaši izveidotā programmā tās salīdzina un nolemj, kurš ir visatbilstošākais variants. Sākotnēji tas tika darīts, lai, panākot anotētāju vienprātību, vienotos par galvenajiem mērķhipotēzes izvirzīšanas principiem, t. i., izstrādātu vadlīnijas, kā teksti šajā korpusā labojami.

VA	**Viņš**	**ir**	**vienpadsmit**	**gadi**			**vinš**	**ir**	**piektais**	**klasē**	**.**
M1	Viņam	ir	vienpadsmit	gadi		,	viņš	ir	piektais	klasē	.
M2	Viņam	ir	vienpadsmit	gadu		,	viņš	ir	piektais	klasē	.
M3	Viņam	ir	vienpadsmit	gadi		,	viņš	ir	piektajā	klasē	.
M4	Viņam	ir	vienpadsmit	gadu		,	viņš	ir	piektajā	klasē	.
M5	Viņš	ir	vienpadsmit	gadus	vecs	,	viņš	ir	piektajā	klasē	.
M6	Viņš	ir	vienpadsmit	gadu	vecs	,	viņš	ir	piektajā	klasē	.

1. tabula. Valodas apguvēja izteikums un iespējamās mērķhipotēzes

1. tabulā redzams, kādas mērķhipotēzes valodas apguvēja rakstītam izteikumam (VA) izvirza trīs tekstu labotāji (anotētāji) (M1–M3), papildus tām tiek piedāvātas vēl dažas iespējamās mērķhipotēzes (M4–M6). Par galīgo mērķhipotēzi izraudzīts M3 variants, jo tādējādi notiek minimāla iejaukšanās (sk. tālāk par mērķhipotēžu izvirzīšanas pamatprincipiem) apguvēja rakstītajā izteikumā: tiek mainīta vietniekvārda *viņš* forma atbilstoši sintaktiskajai struktūrai, lietvārda "gadi" forma netiek mainīta, jo valodas normas pieļauj, ka ar nelokāmu skaitļa vārdu var lietot gan nominatīvu, gan ģenitīvu. Otrreiz lietotajā vietniekvārdā labota pareizrakstības kļūda: *vinš* → *viņš*. Savukārt kārtas skaitļa vārda locījuma forma mainīta, jo no plašāka konteksta var secināt, ka runa ir par piektās klases skolēnu, nevis par piekto skolēnu klasē.

Jau no 1. tabulā aplūkotā piemēra var secināt, ka ir nepieciešami ieteikumi, pēc kuriem valodas apguvēju teksti labojami. Šādi ieteikumi ir vajadzīgi arī tāpēc, lai pētniekiem, kas korpusu pēc tā izveides lietos, būtu saprotams, kā mērķhipotēžu izvirzīšanas principi var ietekmēt iegūtos rezultātus un līdz ar to arī potenciālos secinājumus korpusā veiktajos pētījumos.

2. Pamatprincipi mērķhipotēžu izvirzīšanā

Lai varētu labot korpusa tekstus un vēlāk tos marķēt, ir jānosaka mērķhipotēžu izvirzīšanas pamatprincipi. Veidojot korpusu LaVA, tiek ievēroti četri mērķhipotēžu izvirzīšanas principi, kurus noteikuši korpusa veidotāji.

1. Latviešu literārās valodas normu ievērošana. Lai arī šis princips varētu šķist pašsaprotams, tomēr mācību procesā tas ne vienmēr tiek darīts, piem., ģenitīva lietojums piederības konstrukcijās (1), kas it sevišķi latviešu valodas apguves sākumposmā nereti tiek uzskatīts par mazāk svarīgu.

(1) *man ir draug**i*** → *man nav draug**u***

Tas, vai šo lietojuma niansi studējošajiem mācīt uzreiz vai vēlāk, bieži vien ir attiecīgā docētāja ziņā. Turklāt reizēm tiek izlemts to mācīt tikai daļēji frāzēs (2), (3), bet, neraugoties uz to, korpusā iekļautajos tekstos tādi izteikumi kā *man nav brālis* tiek laboti.

(2) *man nav naudas*

(3) *man nav laika*

Korpusā LaVA iekļauto tekstu autori mācījušies pie daudziem docētājiem, un dažkārt pat vienam un tam pašam studentam dažādos semestros mainījušies arī latviešu valodas docētāji. Atšķiras arī mācību programmas, kursu apjoms un gaidāmie rezultāti. Lai ievērojami nesarežģītu datu ieguvi, korpusam apkopotajos metadatos nav iekļauta informācija par to, cik ilgs ir kurss, kāda ir tā programma, kurš docētājs ir strādājis ar konkrēto studentu un kāda ir bijusi attiecīgā docētāja attieksme pret noteiktiem latviešu valodas normu aspektiem. Turklāt datu ieguves sākumposmā vēl nebija zināms, kādi normu aspekti varētu būt aktuālāki, līdz ar to atšķirības mācīšanas pieejā korpusā pētīt nevarēs. Tāpēc, labojot korpusa LaVA tekstus, konsekvences dēļ nolemts ievērot arī tās latviešu valodas normas, kuras docētāji dažkārt izvēlas ignorēt vai kuru apguve ir iekļauta tikai daļā latviešu valodas kā svešvalodas mācību programmu. Tas attiecas arī uz interpunkciju – lai arī iesācēju līmenī tā vēl nebūtu jāpārzina (Šalme, Auziņa 2016a, 173–174), tās labošana būtiski nemaina citas teksta īpatnības (sk. tālāk minimālās iejaukšanās principu), tāpēc korpusā LaVA nolemts labot pieturzīmju lietojumu, tas arī nākotnē ļautu apguvēju tekstus, piem., automātiski sintaktiski marķēt.

2. Minimāla iejaukšanās. Lai saglabātu apguvēju valodas īpatnības, tiek labots pēc iespējas mazāk. Tāpat jāņem vērā, ka korpusā ir iekļauti teksti tikai no pirmajiem diviem latviešu valodas apguves semestriem. Šajā laikā apguvēju prasmes un zināšanas vēl ir ļoti ierobežotas un nereti neatbilst saturam, ko apguvējs vēlas paust. Bieži vien centienos lietot valodu radoši rodas izteikumi, kas būtu pilnībā jāpārraksta, lai tie atbilstu visām valodas normām (4), (5), tomēr, ja tas tiktu darīts, apguvēju teksti un to labojumi kļūtu tikpat kā nesalīdzināmi, t. sk. arī kļūdu anotējuma aspektā – jo vairāk ir labojumu (kas reizēm pat būtu teksta

pilnīga pārrakstīšana citādi), jo vairāk valodas līmeņos tie tiek veikti, jo sarežģītāka analīze nepieciešama, anotējot kļūdas, un jo grūtāk ir identificēt kļūdas tipu.

(4) *Viņa ir skolotāja kopš divdesmit gadiem*

(5) *Es esmu spēlēt zēnu, miljardieris & filantrops.*

Pārlieku augsta standarta piemērošana būtu pretrunā arī ar mācīšanas un mācīšanās teorijām – pedagoģijā tiek norādīts, ka izaugsme notiek ārpus komforta zonas jeb zināmā robežām, taču pārmērīga attālināšanās no tām noved strauja mācību rezultātu un motivācijas krituma zonā (Luckner, Nadler 1997, 29). Tā kā viens no korpusa izveides mērķiem ir pašpārbaudes uzdevumu veidošana, pārlieku dažādu teksta versiju veidošana, izvirzot mērķhipotēzi, var radīt grūtības automātiski ģenerētajos pašpārbaudes uzdevumos. Līdz ar to korpusā LaVA veikti minimāli labojumi, tiecoties pēc iespējas saglabāt starpvalodas īpatnības, kuras arī var būt nozīmīgs pētījumu objekts (vairāk sk., piem., Selinker 2013, 23–24).

3. Valodas stila saglabāšana sasaucas ar iepriekšējo pamatprincipu – stilistikai latviešu valodas kā dzimtās valodas un latviešu valodas kā otrās valodas apguvē ir ierasts pievērsties pastiprināti, taču pirmajos divos latviešu valodas kā svešvalodas apguves semestros par to vēl ir ļoti maz zināšanu. Ja tiktu laboti visi izteikumi, kas neatbilst latviešu literārās valodas stilistikas normām, mērķhipotēzes tik lielā mērā atšķirtos no sākotnējā teksta, ka to salīdzinājums kļūtu problemātisks (sk. iepriekš tekstā par minimālu iejaukšanos).

4. Netipisko elementu saglabāšana ir cieši saistīta ar valodas stila nelabošanas un minimālas iejaukšanās mērķhipotēzes izvirzīšanas principu – ja stils labots netiek, tiek saglabāts arī netipisks valodas līdzekļu lietojums, ja vien tas nav pret valodas normu. Īpaši bieži tas tiek novērots kā netipiska vārdu secība, piem., partikulas *arī* pievienošana teikuma beigās (6).

(6) *Viņai nepatīk Ukraina un man nepatīk Ukraina **ari**. → Viņai nepatīk Ukraina, un man nepatīk Ukraina **arī**.*

Netipisks lietojums ir starpvalodas īpatnība, kas var būt nozīmīga sintakses apguves un valodu kontaktu pētniekiem, tāpēc to nolemts saglabāt ne vien sākotnējā tekstā, bet arī mērķhipotēzē.

3. Mērķhipotēzes atbilsmes valodas normām

Lai arī četri minētie pamatprincipi LaVA mērķhipotēzēs ir attiecināmi uz visiem valodas līmeņiem, tomēr katrā valodas līmenī ir risināmi problēmgadījumi. Tālāk rakstā mērķhipotēzēs veicamie labojumi raksturoti grupās atbilstoši valodas līmenim, norādot arī biežāk sastopamos kļūdu apakštipus attiecīgajā līmenī.

3.1. Pareizrakstības normas mērķhipotēzēs

Vārdu un vārdformu rakstību, kā arī to, kādi grafiskie simboli izmantojami, atspoguļojot fonēmas, nosaka ortogrāfijas normas. Protams, valodas attīstības gaitā normas mainās, tomēr katrā noteiktā laika posmā pieņemtie ortogrāfijas likumi ir jāievēro. Šie likumi ir jāapgūst ne tikai dzimtās valodas lietotājiem, bet arī tiem, kas mācās valodu kā svešvalodu vai otro valodu.

Latviešu valodā pareizrakstības normas nosaka (Strautiņa, Šulce 2009, 50):

1) morfēmu rakstību vārdos un vārdformās;

2) vārdu rakstību kopā vai šķirti;
3) defisrakstību;
4) lielā sākumburta lietojumu nosaukumos;
5) vārdu pārnesumpārdali;
6) vārdu saīsināšanu;
7) personvārdu rakstību un citvalodu īpašvārdu atveidi.

Izvirzot mērķhipotēzi, tiek labots neatbilstoša burta lietojums, diakritiskās zīmes trūkums vai gluži pretēji – pārdaudzums.

Viena no biežāk sastopamajām atkāpēm no pareizrakstības normām ir neatbilstošas diakritiskās zīmes vai burta lietojums, piem., līdzskaņi tiek rakstīti bez diakritiskajām zīmēm (7), (8), lietots patskaņu kvantitātei neatbilstošs burts (9), (10), kļūdaini rakstīti divskaņi (11), (12):

(7) *vīrie**s**i → vīrie**š**i*
(8) *mēne**s**a → mē**n**eša*
(9) ***a**tri → ā**t**ri*
(10) *v**e**los → v**ē**los*
(11) *na**u** → na**v***
(12) *la**j** → la**i***

Reizēm valodas apguvēji, rakstot vārdus, nejauši vai nezināšanas dēļ izlaiž burtus (13), (14), ieraksta liekus burtus (15), (16) vai pārstata blakus esošos burtus (17), (18). Protams, tas tiek labots.

(13) *kau → kau**t***
(14) *vis → vis**s***
(15) *ko**ll**ekcionēt → ko**l**ekcionēt*
(16) *gru**pp**as → gru**p**as*
(17) *nev**ei**na → nev**ie**na*
(18) *ārt**s**i → ārs**t**i*

Tekstos tiek labota arī valodas normām neatbilstoša vārdu rakstība kopā (19) vai šķirti (20):

(19) *divdesmitdivi → divdesmit divi*
(20) *pus desmitos → pusdesmitos*

Dzimtās valodas vai starpniekvalodas ietekmē valodas apguvēju tekstos ik pa laikam tiek lietoti latviešu valodas alfabētā neesoši burti (21), arī burti ar diakritisko zīmi, kāda latviešu valodas grafētikā nav sastopama (22), (23) (sk. plašāk Kaija 2020).

(21) *Ale**x**ander → Ale**k**sanders*
(22) *mōls → m**o**ls*
(23) *Viņ**n**a ir apkopēja un strādā Radisson Blue viesnīcā. → Viņa ir apkopēja un strādā Radisson Blue viesnīcā.*

Ārvalstu uzņēmumu un organizāciju nosaukumi, arī Latvijas Republikas Uzņēmumu reģistrā reģistrēto uzņēmumu nosaukumi, kas ir svešvalodā vai ir veidoti kā logotipi un kuros bieži vien ir sastopami latviešu alfabētā neesoši burti,

apguvēju tekstos netiek laboti vai pārveidoti, piemēram, *MyFitness, Rimi Express, Stockpot, Maxima.*

Valodas apguvēju tekstos tiek labots arī neatbilstošs lielo un mazo burtu lietojums. Dažkārt valodas apguvēji nelieto lielo sākumburtu teikuma sākumā, reizēm sugasvārdi tiek rakstīti ar lielo burtu citu valodu ietekmē, piem., vācu valodas ietekmē sugasvārdi, savukārt angļu valodas ietekmē valodu, mēnešu nosaukumi u. c. sugasvārdi, kas angļu valodā rakstāmi ar lielo sākumburtu (24).

(24) *Man patīk Eiropieši → Man patīk **e**iropieši*

Lai gan valodas apguvējiem pamatlīmenī vēl netiek mācīts lielo sākumburtu lietojums nosaukumos, izvirzot mērķhipotēzi, nosaukumu rakstība tiek labota atbilstoši latviešu valodas normām (Laugale, Šulce 2012).

Ja vien tas ir iespējams, valodas apguvēju tekstos lietotie citvalodu īpašvārdi tiek rakstīti atbilstoši īpaši noteiktiem atveides principiem (sk., piem., Raģe 1960; Bankava 2004; Placinska 2015; personvarduatveide.lv) un atbilstoši latviešu valodas pareizrakstības noteikumiem (sk. „Noteikumi par personvārdu rakstību un lietošanu latviešu valodā, kā arī to identifikāciju", pieejams: https://likumi.lv/ta/id/85209-noteikumi-par-personvardu-rakstibu-un-lietosanu-latviesu-valoda-ka-ari-to-identifikaciju). Plaši pazīstamu un valodā lietotu īpašvārdu (25)–(27), arī mazāk zināmu īpašvārdu (28), (29) atveide grūtības nerada. Gadījumos, kuros sākotnējā tekstā nav saprotams, no kādas valodas īpašvārds būtu atveidojams un kā to izrunā oriģinālvalodā (30), (31), īpašvārds tiek atstāts oriģinālajā veidolā, figūriekavās, lai varētu vēlāk šādus piemērus izgūt un atbilstoši apstrādāt.

(25) *Germany* → Vācija
(26) *New York* → *Ņujorka*
(27) *Elizabeth* → *Elizabete*
(28) *Essen* → *Esene*
(29) *Lüneburg* → Līneburga
(30) {*Talayhar*}
(31) {*Qamiar*}

3.2. Interpunkcijas normas mērķhipotēzēs

Pieturzīmju lietošana jeb interpunkcija palīdz lasītājam iespējami pilnīgāk uztvert uzrakstītā teksta saturu, kā arī tā atspoguļo teikuma gramatisko struktūru, parāda teksta un teikuma daļu gramatisko saistījumu un dalījumu (Blinkena 2009, 8–9). Lai arī latviešu valodā ir daudz pieturzīmju, piem., punkts, komats, izsaukuma zīme, jautājuma zīme, defise u. c., tomēr „dažādu konstrukciju atdalīšanai un izdalīšanai teikumā visbiežāk tiek lietots komats un domuzīme, retāk – semikols, kols, daudzpunkte, iekavas un pēdiņas" (Šalme, Auziņa 2016b, 32). Arī apguvēju tekstos visas pieturzīmes nav sastopamas.

Jau pamatlīmenī jeb iesācēja līmenī (A1, A2) valodas apguvējam jāprot pareizi izmantot pieturzīmes apgūtajās sintaktiskajās konstrukcijās (Šalme, Auziņa 2016a, 257) – pieturzīmes (punkts, izsaukuma zīme, jautājuma zīme) teikuma beigās, punkts aiz kārtas skaitļa vārdiem, komats, atdalot uzrunu vai uzrunas grupu un partikulas *jā, nē* (Šalme, Auziņa 2016b, 33–34). Tikai vidējā līmenī (B1, B2) un augstākajā līmenī (C1, C2) valodas apguvēji mācās lietot pieturzīmes sarežģītākās

sintaktiskās konstrukcijās, piem., pieturzīmes saliktu teikumu beigās, ar komatu atdala vienlīdzīgus teikuma locekļus, ja starp tiem nav vienojuma saikļa, ar komatu atdala salikta teikuma sastāvdaļas, divdabja teicienus, savrupinājumus (Šalme, Auziņa 2016b, 33–34).

Korpusā LaVA iekļauto tekstu autoriem par latviešu valodas interpunkciju zināšanu vēl nav daudz, arī paši izteikumi studentu tekstos ir diezgan vienkārši, proti, parasti tos veido vienkārši paplašināti vai salikti sakārtoti teikumi, kuros nereti ir daudz vienlīdzīgu teikuma locekļu.

Lai gan zemākajos valodas prasmes līmeņos interpunkcija valodas apguvējam vēl nav jāpārzina pilnībā, izvirzot mērķhipotēzi, tiek labots arī kļūdains pieturzīmju lietojums visos korpusa tekstos:

1) izlaista pieturzīme;
2) lieka pieturzīme;
3) neatbilstoša pieturzīme.

Izlaista pieturzīme

Nereti valodas apguvēji nelieto pieturzīmes tur, kur tām pēc latviešu valodas pareizrakstības normām būtu jābūt – gan teikuma vidū, gan teikuma beigās. Visbiežāk, izvirzot mērķhipotēzi, tiek ielikts izlaists komats, retāk – punkts vai kāda cita pieturzīme.

Apguvēju tekstos komats netiek lietots gan pirms saikļa *un* (32), (33), gan pirms saikļa *bet* (34), kas atdala salikta sakārtota teikuma daļas. Citi sakārtojuma saikļi salikta sakārtota teikuma daļu saistīšanai valodas apguvēju darbos tiek izmantoti reti.

(32) *Es uzvilku kurpes un es ņemu mana jaka. → Es uzvelku kurpes**,** un es ņemu manu jaku.*

(33) *Es studeju medicīna un es dzivoju Rigā. → Es studēju medicīnu**,** un es dzīvoju Rīgā.*

(34) [..] *man patik latvišu valoda un psiholoģija bet man nepatīk histologija* [..] → [..] *man patīk latviešu valoda un psiholoģija**,** bet man nepatīk histoloģija* [..]

Komati netiek lietoti arī vienlīdzīgu teikumu locekļu atdalīšanai (35), bieži vien vārdrindās (36), kas ir bieži sastopamas studentu darbos, viņiem aprakstot, piem., produktus, kas garšo, nosaucot apgūstamos mācību priekšmetus, raksturojot savu ikdienu u. tml.

(35) *Brīevdienās līdoju uz Vāciju apmekleju mana gimēne. → Brīvdienās lidoju uz Vāciju**,** apmeklēju manu ģimeni.*

(36) *Parasti es daru treniņš uz sporta zali „my Fitness” pirmdienā trešdienā un svetdienā. → Parasti es trenējos sporta zālē „My Fitness” pirmdienā**,** trešdienā un svētdienā.*

Komats netiek lietots, atdalot palīgteikumu no virsteikuma (37), (38) un (39), kaut jāatzīst, ka šādas teikumu konstrukcijas A līmeņa darbos ir sastopamas salīdzinoši reti.

(37) *Mans lielakais problems ir ka es nav ēdu pietiekami. → Mana lielākā problēma ir**,** ka es neēdu pietiekami.*

(38) *Vakarā es iešu uz picēriju tāpēc kā man ir ballīte ar Itāliajs asociācija.* → *Vakarā es iešu uz picēriju**,** tāpēc ka man ir ballīte ar Itālijas asociāciju.*

(39) *Ja man nav kolokvijs tad es parasti pavadu laiku ar draugiem.* → *Ja man nav kolokvija**,** tad es parasti pavadu laiku ar draugiem.*

Pieturzīmes netiek lietotas arī iespraudumos (40), savrupinājumos (41). Reizēm iespraudumi tiek atdalīti ar pieturzīmi no vienas puses (40), bet no otras puses ne.

(40) [..] *un vakarā mes ejam uz krogu vai nakskubuu, piemerām Folk klubs ala pagrabs.* > [..] *un vakarā mēs ejam uz krogu vai naktsklubu, piemēram**,** folkkluba Ala pagrabu.*

(41) *Dažreiz es gatavoju kopā ar Hannah mana flatmate.* > *Dažreiz es gatavoju kopā ar Hannu**,** manu dzīvokļa biedreni.*

Izvirzot mērķhipotēzi, tiek novērsts pieturzīmes trūkums teikuma beigās. Valodas apguvēji, īpaši A prasmes līmeņa darbos, nereti aizmirst ielikt pieturzīmi teikuma beigās (42), (43), (44).

(42) *Astoņos vakarā es eju uz sporta klubs* → *Astoņos vakarā es eju uz sporta klubu**.***

(43) *Brokastīs man garšo maizi ar nutellu vai marmeladu* > *Brokastīs man garšo maize ar nutellu vai marmelādi**.***

(44) *Mana ģimere spēlē golfu, es speleju ar viņiem* → *Mana ģimene spēlē golfu, es spēlēju ar viņiem**.***

Nākas labot arī pēdiņu lietojumu. Latviešu valodā pēdiņās raksta dažādus nosaukumus, piem., mākslas darbu, uzņēmumu, dažādu mašīnu un ierīču, ēdienu, dzērienu nosaukumus, ordeņu u. c. nosaukumus (Blinkena 2009, 398–401), tomēr pēdējā laikā latviešu literārajā valodā pēdiņas vai kāds cits burtveidols (kursīvs), lai atdalītu nosaukumu no pārējiem izteikuma vārdiem, lietots netiek. Tāpēc arī mērķhipotēzēs pēdiņas tiek lietotas tikai tad, ja tās ir lietojis teksta autors (45), ja teksta autors pēdiņas nav lietojis, tas labots netiek (46). Ja autors ir lietojis vienpēdiņas, arī tas tiek uzskatīts par pēdiņu lietojumu, taču tiek labots uz pēdiņām.

(45) *Es eju uz pieturu **„Elizabetes iela”**.*

(46) *Rīgā es sportoju sporta klubā, **MyFitness**.*

Liekas pieturzīmes (pieturzīmju pārdaudzums)

Lai gan biežāk valodas apguvēju tekstos vērojamas izlaistas pieturzīmes, tomēr ir gadījumi, kuros pieturzīmes lietotas nevietā. Visbiežāk liekas pieturzīmes valodas apguvēji lieto, 1) atdalot situantu no pārējā teikuma, 2) aiz virsraksta.

Lai arī latviešu valodā galvenais ir gramatiskais interpunkcijas princips, proti, ar pieturzīmēm rādīts teksta gramatiskais saistījums un dalījums, atklājot teksta gramatisko struktūru, tomēr nozīmīgs ir arī jēdzieniskais interpunkcijas princips (Blinkena 2009, 24). Tomēr ir vairāki gadījumi, kad sintagmatiskais dalījums neatbilst strukturāli gramatiskajam dalījumam, tātad veidojas pauze, kura rakstos nav apzīmējama ar pieturzīmi, proti, pauze bieži veidojas aiz teikuma sākumā nostatītas izvērstas apstākļu grupas (Blinkena 2009, 24–25). Iespējams, pauzes dēļ vai arī pēc citu valodu parauga valodas apguvēji samērā bieži aiz situanta teikuma sākumā (47), (48) liek komatu.

(47) *Pudienās, es normali eju sviestmaizi. → Pusdienās es parasti ēdu sviestmaizi.*

(48) *Nedeļas nogalē, es celos deviņos. → Nedēļas nogalē es ceļos deviņos.*

Daļa apguvēju tekstam ir pievienojuši arī virsrakstu. Reizēm aiz tā tiek lietots punkts (49), bet biežāk punkta nav. Tādēļ korpusa veidotāji ir vienojušies, labojot valodas apguvēju tekstus, punktu aiz nosaukuma nelikt. To pieļauj arī latviešu valodas pareizrakstības normas, kas nosaka, ka aiz virsrakstiem īpaša pieturzīme, kas atdalītu virsraksta teikumu no pārējā teksta, nav vajadzīga. Pieturzīme aiz virsraksta ir liekama tikai tad, ja tā bez teikuma pabeigtības rāda arī teikuma modalitāti (Blinkena 2009, 154).

(49) *Mana diena**.*** *→ Mana diena*

Protams, izvirzot mērķhipotēzi, tiek labotas arī pārrakstīšanās kļūdas, piem., teikuma vidū lietotais punkts tiek dzēsts (50).

(50) *Mans patik slidot. un zīmēt. → Mans patīk slidot un zīmēt.*

Izvirzot mērķhipotēzi, tiek labota kļūdaina pieturzīmes atrašanās vieta – reizēm valodas apguvēju darbos komats tiek likts nevis pirms saistītājvārda, bet aiz tā (51).

(51) *Es ēdu brokastis bet**,*** *man ir ļoti mazs brokastis. → Es ēdu brokastis**,*** *bet man ir ļoti mazas brokastis.*

Neatbilstoša pieturzīme

„Teikuma beigu pieturzīmju lietojuma noteikumi latviešu valodā lielākoties ir tādi paši kā citās Eiropas valodās. Punkts, izsaukuma zīme, jautājuma zīme rāda ne tikai teikuma beigas, bet arī pauž valodas lietotāja attieksmi pret izteikto saturu." (Šalme, Auziņa 2016b, 32) Reizēm ir lietota neatbilstoša pieturzīme, piem., jautājuma teikuma vai izsaukuma teikuma (52) beigās likts punkts, tāpat arī tiek lietota jautājuma zīme teikuma beigās, kur tā nebūtu lietojama. Lai saprastu, vai tiek lietota neatbilstoša pieturzīme teikuma beigās, nereti ir nepieciešams plašāks konteksts.

(52) *Visu labu**.*** *→ Visu labu**!***

3.3. Formveidošanas un vārddarināšanas normas mērķhipotēzēs

Korpusa tekstos ļoti bieži ir sastopams neatbilstošu vārdformu lietojums. Valodas apguvē prasme izveidot noteiktu vārdformu nenodrošina prasmi to lietot (par to vairāk sk. Laizāne 2012, 2013). Bieži vien, pat ja konkrētā forma (iespējams, ar labojamām pareizrakstības kļūdām) ir vārda paradigmā, tā nav lietojama attiecīgajā konstrukcijā. Tādas ir, piem., neatbilstošas darbības vārdu personas formas (53), neatgriezeniskā darbības vārda formas lietojums atgriezeniskā darbības vārda vietā (54), nenoteiktās galotnes lietojums noteiktās galotnes vietā (55) u. tml.

(53) *Šobrid es **dzivo** Rigā. → Šobrīd es **dzīvoju** Rīgā.*

(54) *Es parasti **pamodināju** septīnos no rīta. → Es parasti **pamostos** septiņos no rīta.*

(55) *Mani **vecāks** bralis ir Daniel un mani **jaunāks** brali ir Leo. → Mans **vecākais** brālis ir Daniels, un mans **jaunākais** brālis ir Leo.*

Novirzes nereti mēdz būt lietvārdu, vietniekvārdu, skaitļa vārdu, īpašības vārdu locījumu lietojumā, piem., lietojot nominatīvu ģenitīva vietā (56), (57) vai akuzatīva vietā (58), (59).

(56) *Man ir daudz **skolotajas*** [..] → *Man ir daudz **skolotāju*** [..]

(57) [..] *piektdienā man nav **lekcijas**.* → [..] *piektdienā man nav **lekciju**.*

(58) *Es parasti pērku **jogurts*** [..] → *Es parasti pērku **jogurtu*** [..]

(59) ***Viņa** sauc Lēna* [..] → ***Viņu** sauc Lēna* [..]

Iespējams, daļa no novirzēm, kas tiek labotas, izvirzot mērķhipotēzes, valodas apguves procesā (vismaz A līmenī) par tādām netiek uzskatītas – piem., (57) piemērā ģenitīva lietojumu nosaka darbības vārds *nav*, un „valodas praksē šajās konstrukcijās nereti lieto arī nominatīvu" (Smiltniece 2013, 349; sk. arī Kalnača 2014, 54; Kalnača, Lokmane, Metslang 2019, 60–62). Piem., latviešu valodas mācību grāmatā zobārstniecības studentiem šādā konstrukcijā ģenitīva lietojums tiek aplūkots tikai vienā no pēdējām nodaļām (Laizāne, Kaija 2020). Tomēr korpusā LaVA šāds lietojums tiek labots.

Vietām var rasties domstarpības par to, vai noteikta kļūda ir drīzāk pareizrakstības vai morfoloģijas kļūda, ja atšķirība starp sākotnējo tekstu un mērķhipotēzi ir, piem., vienas diakritiskās zīmes vai viena burta lietojumā (60). Tomēr mērķhipotēzes izvirzīšanā tam nav lielas lomas – šis jautājums ir apspriežams kļūdu klasifikācijas, nevis labošanas stadijā, un labojuma forma nav atkarīga no kļūdas iespējamās klasifikācijas.

(60) *Es ari reti studēju darba dienā un studeju daudz nedēļas **nogale**.* → *Es arī reti studēju darba dienā un studēju daudz nedēļas **nogalē**.*

Dažkārt lietots neatbilstošas dzimtes lietvārds (61), (62), īpašības vārds (63), skaitļa vārds (64), vietniekvārds (65). Ja izteikumā ir informācija, kas ļauj saprast, kura dzimte būtu lietojama, visi vārdi, kas tai neatbilst, mērķhipotēzē tiek attiecīgi laboti.

(61) [..] *mas tevs ir **architecte*** [..] → [..] *mans tēvs ir **arhitekts*** [..]

(62) *Mans tēvs ir pediatrs un mana māte ir **terapeits**.* → *Mans tēvs ir pediatrs, un mana māte ir **terapeite**.*

(63) *Es dzīvoju centra, mans dzīvokli ir **liela**.* → *Es dzīvoju centrā, mans dzīvoklis ir **liels**.*

(64) *Mana māte ir četrdesmit **astonas** gads* [..] → *Manai mātei ir četrdesmit **astoņi** gadi* [..]

(65) *Man ir trīs māsas.* [..] ***Viniēm** garšo kafija.* → *Man ir trīs māsas.* [..] ***Viņām** garšo kafija.*

Ja dzimte ir atkarīga no saskaņojuma ar tā apkaimē esošo pareizi lietotu vārdu, tad tā attiecīgi tiek labota. Tomēr ne vienmēr pietiek ar vārda tiešo apkaimi, lai konstatētu šādas neatbilstības, it sevišķi, ja runa ir par cilvēkiem, dažkārt – arī mājdzīvniekiem. Korpusā iekļautos tekstus reizēm nākas skatīt kopumā, lai saprastu, vai tajā par vienu un to pašu personu runāts konsekventā dzimtē (65).

Gadījumos, kuros leksēmas, piemēram, *tēvs, māte, māsa,* izvēle norāda arī uz personas dzimumu (61), (62), (65), labojumi tiek veikti atbilstoši tam. Dzimumu dažkārt var noteikt arī pēc personvārda (66). Citkārt vienīgā informācija par aprakstāmās personas dzimumu var būt gramatiskās dzimtes lietojums, un,

ja dzimte tekstā ir lietota nekonsekventi, ne vienmēr ir skaidrs, kurā dzimtē to vienādot (67), (68). Šādos gadījumos mēģināts noteikt, kura dzimte tekstā dominē, un pieskaņot pārējās teksta daļas tai, taču, protams, šādos gadījumos ir iespējams, ka labotāja interpretācija nav pilnīgi pareiza.

(66) *Mani sauc Jānis, un es esmu **skoltāja**. → Mani sauc Jānis, un es esmu **skolotājs**.*

(67) *Es esmo Waface.* [..] *Es esmo medicinas **stodonte**.* [..] *Riga esmo **uztorejies** jau 4 menesus. → Es esmu Vafase.* [..] *Es esmu medicīnas **studente**.* [..] *Rīgā esmu **uzturējusies** jau 4 mēnešus.*

(68) *Man ir anglu valoda ari, un anglu valoda ir arī interesanta, jo mums ir **professors** no Amerikas. Vina ir ļoti laba **profesors**. → Man ir angļu valoda arī, un angļu valoda ir arī interesanta, jo mums ir **profesore** no Amerikas. Viņa ir ļoti laba **profesore**.*

Vērojamas arī grūtības lietvārdu skaitļa lietojumā (69), (70), (71). Ļoti bieži lietotas formas, kas latviešu valodas gramatikai neatbilst neatkarīgi no konteksta, tomēr ir gadījumi, kuros papildus jāņem vērā arī semantika. Piem., konstrukcijās ar *garšo* kā teikuma priekšmets var tikt lietots lietvārds gan vienskaitlī, gan daudzskaitlī, taču atbilstoši latviešu valodas gramatikai daudzskaitlim piemīt vairāku priekšmetu vai vielu kopuma, kategorijas nozīme (Smiltniece 2013, 337) (70), (71), bet vienskaitlis tiek lietots tad, ja runa ir par konkrētu vienu priekšmetu vai par vielu (Smiltniece 2013, 339) (72).

(69) *Man ir divdesmit viens **gadi**. → Man ir divdesmit viens **gads**.*

(70) *Man garso **auglis**. → Man garšo **augļi**.*

(71) *Pusdienas es edu kartupelus man loti gārso **kartupelis**. → Pusdienās es ēdu kartupeļus, man ļoti garšo **kartupeļi**.*

(72) *Vins garšo **piens** un **baltmaize**. → Viņam garšo **piens** un **baltmaize**.*

Tāpat ir arī gadījumi, kuros veidotas latviešu valodā neesošas formas. Šādi piemēri konstatēti dažādu lokāmu vārdšķiru vārdiem: lietvārdiem (73), darbības vārdiem (74), skaitļa vārdiem (75), kā arī īpašības vārdiem (76).

(73) *Es esmu otrajā **semestrā*** [..] *→ Es esmu otrajā **semestrī*** [..]

(74) *Es pusdienu universitātē ar maniem draugiem un mēs runājam un mēs **smietiem** daudz. → Es pusdienoju universitātē ar maniem draugiem, un mēs runājam, un mēs **smejamies** daudz.*

(75) *Divi tūkstoši četrajā gadā* [..] *→ Divi tūkstoši **ceturtajā** gadā* [..]

(76) [..] *man nepatik **aukst** laiks! →* [..] *man nepatīk **auksts** laiks.*

Sākot apgūt gramatikas likumus, izņēmumi vai nesistēmiskums var radīt grūtības, un nereti tiek veidotas gramatizētas nelokāmo vārdu formas (77), it īpaši uzņēmumu nosaukumos (78). Lai arī šādu formu lietojums liecina par zināmu gramatikas (piem., lokatīva) nozīmes izpratni, turklāt ir izplatīts sarunvalodā, tās tomēr tiek labotas atbilstoši latviešu literārās valodas normām, proti – pārveidotas nelokāmā formā.

(77) *Jums ir daudziem ļotiem labiem produktiem. → Jums ir daudz ļoti labu produktu.*

(78) *Biezi es edu pusdiena universitatē vai **stockpodā**. → Bieži es ēdu pusdienas universitātē vai **Stockpot**.*

Apgūstot lietvārda formu veidošanu, apguvējam diezgan agri nākas saskarties arī ar līdzskaņu miju, it īpaši – vēsturisko līdzskaņa *j* noteikto līdzskaņu miju, kas parādās, piem., 2. deklinācijas lietvārdu vienskaitļa ģenitīva un daudzskaitļa formās (vairāk sk. Kalnača 2004, 71; Auziņa 2013, 91–92; Kalnača 2013, 161–165; Kalnača 2014, 9–11). Lai arī mācību līdzekļos tā parasti tiek skaidrota (piem., Klēvere-Velhli, Naua 2012, 44; Auziņa, Berķe u. c. 2016, 159), valodas apguves sākumposmā šis bieži vien tiek uzskatīts par mazāk nozīmīgu jautājumu, tāpēc likumsakarīgi, ka arī korpusā iekļautajos tekstos formas bez līdzskaņu mijas parādās un tiek atbilstoši labotas (79), (80).

(79) [..] *un dzieru tēju* ***zālu***. → [..] *un dzeru tēju* ***zāļu***.

(80) *Man ļoti labi garšo rīsi ar* ***dārzeniem***. → *Man ļoti labi garšo rīsi ar* ***dārzeņiem***.

Valodas apguvēju tekstos tiek labots kļūdains prievārdu saistījums ar lietvārdiem – gan ar neiederīgām, gan ar neeksistējošām formām (81), (82), (83), (84).

(81) *Tas ir pie* ***Šveici*** *un* ***Franciji*** → *Tas ir pie* ***Šveices*** *un* ***Francijas***.

(82) *Es dzivoju kopa ar* ***divas meitenes*** → *Es dzīvoju kopā ar* ***divām meitenēm***.

(83) *Mani sauc Peter un es esmu no* ***spanijā***. → *Mani sauc Pēters, un es esmu no* ***Spānijas***.

(84) *Es dejoju tango kopa ar* ***mana draudzene***. → *Es dejoju tango kopā ar* ***manu draudzeni***.

Morfoloģijas un vārddarināšanas normu ievērošana mērķhipotēzēs lielākoties nesagādā grūtības, ja vien ir skaidrs, ko autors tekstā ir gribējis pateikt. Tomēr teksta izpratni dažkārt mēdz apgrūtināt arī novirzes no normas sintakses un leksikas līmenī.

3.4. Leksikas normas mērķhipotēzēs

Izvirzot leksikas lietojuma mērķhipotēzes, labojumu ir salīdzinoši maz. Tiek ievērots minimālās iejaukšanās princips, turklāt arī valodas stils netiek labots, tādēļ atkāpes no normas, kas dzimtās valodas runātāju tekstos tiktu labotas, apguvēju tekstos labotas netiek.

Reizēm apguvēju darbos vārdi netiek lietoti atbilstošā nozīmē (85)–(89), tāpēc, izvirzot leksikas lietojuma mērķhipotēzi, tiek vērtēta vārdu kontekstuālā iederība, īpaši – semantiskā atbilsme. Tiek analizēts izteikuma saturs, un katrā semantisko pārviržu (arī nelielu) gadījumā valodas apguvēja lietotais vārds tiek aizstāts ar kontekstuāli iederīgu vārdu (85–89).

(85) [..] *bet pārsvarā es ēdu brokastu* ***labību***. → [..] *bet pārsvarā es ēdu brokastu* ***pārslas***.

(86) *Turklat es* ***daru*** *bizi.* → *Turklāt es* ***pinu*** *bizi.*

(87) *Mes iesim uz kino rīt* ***redzēt*** *„Avengers – Endgame”.* → *Mēs iesim uz kino rīt* ***skatīties*** *„Avengers – Endgame”.*

(88) *Vinem garšo šokolāde, augli un saldejums bet* ***nav*** *tomati.* → *Viņiem garšo šokolāde, augļi un saldējums, bet* ***negaršo*** *tomāti.*

(89) *Viņa strādā kafejnīcā Vecrigā un iet ar pēda uz darbā.* → *Viņa strādā kafejnīcā Vecrīgā un iet ar* ***kājām*** *uz darbu.*

Tiek aizstāti arī tie vārdi (91–93) vai vārdu savienojumi (94), kuriem citu valodu (visbiežāk angļu vai krievu valodas) ietekmē klāt nākušas papildu nozīmes, vai tie tiek lietoti ar kādu nozīmes papildkomponentu. Piemēram, ļoti bieži valodas apguvēju tekstos, runājot par ēdieniem vai dzērieniem, tiek lietots darbības vārds *mīlēt* ar nozīmi 'būt patikai (pret ko); garšot'. Izvirzot mērķhipotēzi, šādos izteikumos vārds *mīlēt* tiek aizstāts ar vārdu *garšot* vai *patikt* (93).

(90) *sērija → seriāls*

(91) *Bet es ienīstu **klasi** pēcpusdienā. → Bet es ienīstu **nodarbību** pēcpusdienā.*

(92) [..] *es **mīlu** Ziemassvētku brīvdienas.* → [..] *man **patīk** Ziemassvētku brīvdienas.*

(93) *Es **milu** dzert kafija bez cukura. → Man **patīk** dzert kafiju bez cukura.*

(94) [..] *un es kļūšu par labu **biznesa cilvēku**.* → [..] *un es kļūšu par labu **uzņēmēju**.*

Konstatēts arī neatbilstošu priedēkļverbu lietojums (95), (96).

(95) ***aiz**braukt → **no**braukt*

(96) ***Uz**gaidiešu atbilde! → Gaidīšu atbildi!*

Valodas apguvēju darbos vērojamas nepilnības prievārdu lietojumā: 1) tiek izvēlēts neprecīzs prievārds, 2) reizēm prievārds tiek lietots latviešu valodai netipiskā nozīmē. Prepozicionālos vārdu savienojumus, kur, mainot lietojamo prievārdu, var mainīties arī locījums, kādā būtu lietojami ar to saistītie vārdi, rodas sekundārās kļūdas (par sekundārajām kļūdām vairāk sk. Kaija 2019).

Sākot mācīties valodu, tekstos „iesprūk" kāds valodas apguvēja dzimtās valodas vai starpniekvalodas vārds. Ik pa laikam lieto vārdus svešvalodās, piem., *und* 'un', *but* 'bet' (97), *is* 'ir', *et* 'un'.

(97) *Man ļoti patīk cilvēki **but** nepatīk pilsētas. → Man ļoti patīk cilvēki, **bet** nepatīk pilsētas.*

3.5. Sintakses normas mērķhipotēzēs

Par mērķhipotēzi sintaksē var runāt ļoti nosacīti. Faktiski, izvirzot mērķhipotēzi, valodas apguvēja teksts tiek sagatavots tālākai analīzei, t. i., kļūdu marķēšanai. Galvenie labojumi, kas tiek veikti valodas apguvēju izteikumos, ir šādi:

1) tiek dzēsts lieks vārds;

2) sintaktiskā konstrukcija tiek papildināta, ierakstot tajā izlaistu vārdu.

Latviešu valodai netipiska vārdu secība, labojot tekstus, netiek mainīta.

Nereti valodas apguvēji lieto prievārda konstrukcijas locījuma formas vietā. Labojot tekstus, prievārda konstrukcija, ko veido lietvārds vai vietniekvārds kopā ar prievārdu, tiek aizstāta ar atbilstošu locījuma formu (98–101).

(98) *bieži man nav gana laika **par** brokastis* [..] → *bieži man nav gana laika brokastīm* [..]

(99) *Kad es esmu māja, es gatavoju vakariņas **par** mani un mans vīrs. → Kad es esmu mājās, es gatavoju vakariņas man un manam vīram.*

(100) *Dažreiz es dejoju **uz** deju studijā. → Dažreiz es dejoju deju studijā.*

(101) ***No** vakar mes lasiem bērnu grāmatas. → Vakarā mēs lasām bērnu grāmatas.*

Reizēm valodas apguvēji rīkojas pretēji, proti, tiek lietota locījuma forma prievārda konstrukcijas vietā (102), (103).

(102) *Desmitos es eju **pieturā** ar draugiem. → Desmitos es eju **uz pieturu** ar draugiem.*

(103) *Es braucu **universitātē**. → Es braucu **uz** universitāti.*

Tāpat nereti tiek lietots darbības vārds kopā ar lietvārdu akuzatīvā, kur latviešu valodā ir attiecīgas nozīmes darbības vārds. Domājams, tas notiek valodas apguvēju dzimtās valodas vai starpniekvalodas (parasti angļu valodas) ietekmē (104), (105).

(104) [..] *izbraucu kopā ar draugiem vai **spēlēšu sportu**. →* [..] *izbraucu kopā ar draugiem vai **sportoju**.*

(105) *Man patīk **darīt sportu** un dejot. → Man patīk **sportot** un dejot.*

Latviešu valodā, tāpat kā citās valodās, adverbiālā vai nominālā izteicējā saitiņas funkcijā lietotais verbs *būt* var būt izlaists (Kalnača 2013, 467), tomēr ir konstrukcijas, kur tas ir nepieciešams. Tādos gadījumos izteikums tiek papildināts ar saitiņu (106), (107).

(106) *Es nevāru, jo vārīt sarežģīti* [..] *→ Es nevāru, jo vārīt **ir** sarežģīti* [..]

(107) *Viņš divi gadi. → Viņam **ir** divi gadi.*

Reizēm valodas apguvēji izlaiž finīto darbības vārdu saliktā verbālā izteicējā (108).

(108) *Inese * gulēt divpadsmitos. → Inese **iet gulēt** divpadsmitos.*

Secinājumi

Kļūdu marķējums ir valodas apguvēju korpusa galvenā iezīme. Savukārt kļūdu identifikācija vienmēr balstās uz noteiktā veidā rekonstruētu valodas apguvēja izteikumu (mērķhipotēzi). Lai korpusā LaVA mērķhipotēzes tiktu izvirzītas konsekventi, ir izveidotas vadlīnijas, kā labot un rekonstruēt valodas apguvēja tekstus. Rakstā ir izklāstīti galvenie mērķhipotēžu izvirzīšanas principi. Pamatā labojumi tiek veikti, ievērojot latviešu valodas normas. Tomēr ir gadījumi, kad, izvirzot mērķhipotēzi, atkāpes no normas ir pieļaujamas. Darba gaitā vadlīnijas var tikt atjauninātas un pilnveidotas, ietverot līdz tam neaplūkotus gadījumus.

Izvirzot mērķhipotēzi, ir skaidrs, ka tā var aptvert tikai noteiktu lingvistiskās informācijas daudzumu. Ne vienmēr ir viegli nosakāms, vai labojums saistīts ar formveidošanu un vārddarināšanu vai tomēr sintaksi, jo vārdformu lietojums bieži vien izriet no sintaktiskās konstrukcijas. It sevišķi pamanāms tas ir, piem., vārdu savienojumos ar prievārdu, kur, mainot lietojamo prievārdu, var mainīties arī locījums, kādā būtu lietojami ar to saistītie vārdi, šādi radot sekundāro kļūdu.

Kļūdu tipi ir atkarīgi arī no mērķhipotēzes pamatā esošās koncepcijas, un pašlaik LaVA korpusā ir noteikti šādi kļūdu tipi:

1) pareizrakstības kļūdas;
2) interpunkcijas kļūdas;
3) formveidošanas un vārddarināšanas kļūdas;
4) sintakses kļūdas;

5) leksikas kļūdas;
6) sekundārās kļūdas.

Izvirzītās mērķhipotēzes daļēji atvieglina kļūdu marķēšanu. Tā kā valodas apguvēju teksti tiek laboti, ir iespējams sastatīt oriģinālo un laboto tekstu, kā arī veikt automātisku tekstu morfoloģisko analīzi. Tas ļauj daļu kļūdu, konkrēti – pareizrakstības, interpunkcijas, formveidošanas un vārddarināšanas, kā arī leksikas kļūdas – noteikt automātiski un pēc tam manuāli pārliecināties par kļūdas atbilstību konkrētam tipam.

Saīsinājumi un apzīmējumi

A1 valodas prasmes pamatlīmeņa apakšlīmenis, *Breakthrought*
A2 valodas prasmes pamatlīmeņa apakšlīmenis, *Waystage*
LaVA *Latviešu valodas apguvēju korpuss*
LOC lokatīvs
M mērķhipotēze
NOM nominatīvs
SG vienskaitlis
VA valodas apguvēja izteikums

Avots

Latviešu valodas apguvēju korpuss. Pieejams: http://lava.korpuss.lv

Literatūra

1. Auziņa, Ilze. 2013. Latviešu valodas fonētiski fonoloģiskie procesi un likumi. *Latviešu valodas gramatika.* Nītiņa, Daina, Grigorjevs, Juris (red.). Rīga: LU Latviešu valodas institūts, 80–95.
2. Auziņa, Ilze, Berķe, Maija, Lazareva, Anta, Šalme, Arvils. 2016. *A2 LAIPA. Latviešu valoda. Mācību grāmata*. Rīga: Latviešu valodas aģentūra.
3. Bankava, Baiba (sast.). 2004. *Franču īpašvārdu atveide latviešu valodā*. Bankavs, Andrejs (red.). Rīga: Zinātne.
4. Blinkena, Aina. 2009. *Latviešu interpunkcija.* Rīga: Zvaigzne ABC.
5. Citvalodu personvārdu atveide latviešu valodā. Pieejams: personvarduatveide.lv
6. Dagneaux, Estelle, Denness, Sharon, Granger, Sylviane. 1998. Computer-aided error analysis. *System*. 26 (2), 163–174.
7. Ellis, Rod. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.
8. Gilquin, Gaëtanelle, De Cock, Sylvie, Granger, Sylviane. 2010. *Louvain international database of spoken English interlanguage* (*CD-ROM + handbook*). Louvain-la-Neuve: Presses universitaires de Louvain.
9. Granger, Sylviane, Dagneaux, Estelle, Meunier, Fanny, Paquot, Magali. 2009. *International corpus of learner English v2* (*Handbook + CD-Rom*). Louvain-la-Neuve: Presses universitaires de Louvain.
10. James, Carl. 1998. *Errors in language learning and use.* London: Addison Wesley Longman.

11. Kaija, Inga. 2019. Sekundārās kļūdas valodas apguvēju korpusos. *Valodu apguve: problēmas un perspektīva.* XV, 30–38.
12. Kaija, Inga. 2020. Jaunu burtu veidošana ar diakritiskajām zīmēm latviešu valodas kā svešvalodas apguvēju tekstos. *Valodu apguve: problēmas un perspektīva*. XVI (iespiešanā).
13. Kaija, Inga, Laizāne, Inga. 2020. *Latviešu valoda zobārstniecībā*. Rīga: Rīgas Stradiņa universitāte.
14. Kalnača, Andra. 2004. *Morfēmika un morfonoloģija*. Rīga: LU Akadēmiskais apgāds.
15. Kalnača, Andra. 2013. Morfonoloģija. *Latviešu valodas gramatika.* Nītiņa, Daina, Grigorjevs, Juris (red.). Rīga: LU Latviešu valodas institūts, 154–189.
16. Kalnača, Andra. 2014. *A typological perspective on Latvian grammar.* Warsaw/ Berlin: De Gruyter.
17. Kalnača, Andra, Lokmane, Ilze, Metslang, Helena. 2019. Subject case alternation in Latvian and Estonian existential clauses. *Eesti rakenduslingvistika ühingu aastaraamat / Estonian papers in applied linguistics*. 15, 53–82.
18. Klēvere-Velhli, Inga, Naua, Nikole. 2012. *Latviešu valoda studentiem. Mācību līdzeklis latviešu valodas kā svešvalodas apguvei.* Rīga: Latviešu valodas aģentūra.
19. Laizāne, Inga. 2012. Akuzatīvs latviešu valodas kā svešvalodas apguvē. *Valoda – 2012. Valoda dažādu kultūru kontekstā.* XXII, 194–203.
20. Laizāne, Inga. 2019. *Latviešu valoda kā svešvaloda: lingvodidaktikas virziena attīstība Latvijā un ārpus tās*. Promocijas darbs filoloģijas doktora zinātniskā grāda iegūšanai valodniecības nozarē. Liepāja: Liepājas Universitāte.
21. Laizāne, Inga. 2013. Ģenitīva locījums un tā nozīmju lietojums latviešu valodas kā svešvalodas apguvē. *Valodu apguve: problēmas un perspektīva*. IX, 82–93.
22. Laizāne, Inga, Kaija, Inga. 2020. „Latviešu valoda zobārstniecības studentiem" – pirmā latviešu valodas kā svešvalodas mācību grāmata zobārstniecības studentiem. *Valodu apguve: problēmas un perspektīva*. XVI (iespiešanā).
23. Laugale, Velga, Šulce, Dzintra. 2012. *Lielo burtu lietojums latviešu valodā: ieskats vēsturiskajā izpētē, problēmas un risinājumi*. Rīga: Latviešu valodas aģentūra.
24. Leech, Geoffrey. 1998. Preface. *Learner English on computer*. Granger, Sylviane (ed.). London: Addison Wesley Longman, xiv–xx.
25. Luckner, John L., Nadler, Reldan S. 1997. *Processing the experience: strategies to enhance and generalize learning*. Dubuque: Kendall / Hunt Publishing Company.
26. Lüdeling, Anke, Doolittle, Seanna, Hirschmann, Hagen, Schmidt, Karin, Walter, Maik. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache.* 2, 67–73.
27. Lüdeling, Anke, Walter, Maik, Kroymann, Emil, Adolphs, Peter. 2005. Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*, 15–17.
28. Mendes, Amália, Antunes, Sandra, Janssen, Maarten, Gonçalves, Anabela. 2016. The COPLE2 corpus: a learner corpus for Portuguese. *Proceedings of the 10th edition of the language resources and evaluation conference* (LREC). Calzolari, Nicoletta et al. (eds.). Portorož: European Language Resources Association (ELRA), 3207–3214.
29. Nesselhauf, Nadja. 2005. *Collocations in a learner corpus.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

30. Raģe, Silvija. 1960. *Norādījumi par citvalodu īpašvārdu pareizrakstību un pareizrunu latviešu literārajā valodā.* I. *Igauņu valodas īpašvārdi*. Rīga: Latvijas PSR Zinātņu akadēmijas izdevniecība.
31. Rakhilina, Ekaterina, Vyrenkova, Anastasia, Mustakimova, Elmira, Ladygina, Alina, Smirnov, Ivan. 2016. Building a learner corpus for Russian. *Proceedings of the joint workshop on NLP for computer assisted language learning and NLP for language acquisition at SLTC*. Volodina, Elena, Grigonytė, Gintarė, Pilán, Ildikó, Nilsson Björkenstam, Kristina, Borin, Lars (eds.). Linköping: LiU Electronic Press, 66–75.
32. Reznicek, Marc, Lüdeling, Anke, Hirschmann, Hagen. 2013. Competing target hypotheses in the Falko corpus. *Automatic treatment and analysis of learner corpus data*. Diaz-Negrillo, Ana, Ballier, Nicolas, Thompson, Paul (eds.). Amsterdam: John Benjamins Publishing Company, 101–124.
33. Placinska, Alla. 2015. *Portugāļu īpašvārdu atveide latviešu valodā: ieteikumi*. Rīga: Latviešu valodas aģentūra.
34. Selinker, Larry. 2013. *Rediscovering interlanguage*. London: Routledge.
35. Siemen, Peter, Lüdeling, Anke, Müller, Frank Henrik. 2006. FALKO-ein fehlerannotiertes Lernerkorpus des Deutschen. *Proceedings of KONVENS 2006*. Butt, Miriam (ed.). Konstanz: Universität Konstanz, 130–136.
36. Smiltniece, Gunta. 2013. Lietvārds (substantīvs). *Latviešu valodas gramatika.* Nītiņa, Daina, Grigorjevs, Juris (red.). Rīga: LU Latviešu valodas institūts, 324–369.
37. Strautiņa, Vaira, Šulce, Dzintra. 2009. *Latviešu valodas pareizruna un pareizrakstība.* Rīga: RaKa.
38. Šalme, Arvils, Auziņa, Ilze. 2016a. *Latviešu valodas prasmes līmeņi: pamatlīmenis A1, A2, vidējais līmenis B1, B2.* Rīga: Latviešu valodas aģentūra.
39. Šalme, Arvils, Auziņa, Ilze. 2016b. *Latviešu valodas prasmes līmeņi: augstākais līmenis C1, C2. Vadlīnijas.* Rīga: Latviešu valodas aģentūra.
40. Tono, Yukio. 2003. Learner corpora: design, development and application. *Proceedings of the corpus linguistics 2003 conference*. Archer, Dawn, Rayson, Paul, Wilson, Andrew, McEnery, Tony (eds.). Lancaster: Lancaster University, 800–809.
41. Znotiņa, Inga. 2018. *Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas.* Promocijas darbs filoloģijas doktora zinātniskā grāda iegūšanai valodniecības nozarē. Liepāja: Liepājas Universitāte.

Summary

A learner corpus is a computerized textual database of the language produced by foreign language learners. Such corpus enables researchers to create more efficient learning materials and teaching methodology for language learners by using the corpus-driven error analysis. The learner's corpus, like other language corpora, can be annotated at different language levels (morphologically, syntactically); however, corpus-based error annotation and the corpus-based error analysis are especially important in the learner's language research. Error analysis is influenced by certain factors: 1) the error types setup or error typology; and 2) target hypothesis setup, e. g., corrected text. Therefore, it is crucial to have special guidelines indicating the subject of annotation and the methods how the annotation is performed.

The article begins with description of “The Latvian Learner corpus” (LaVA) and its initial development strategies, the term of target hypothesis and its role in the creation of the learner corpus. The main target hypothesis setup criteria in the LaVa corpus is also provided with the examples showing how the language learners’ utterances are being corrected according to the language norms, and the main deviations from the rules allowed.

This work has received financial support from the Latvian Council of Science under the grant agreement No. lzp-2018/1-0527 (“Development of Learner Corpus of Latvian: methods, tools and applications”) in synergy with the Latvian State Research Programme “Latvian Language”, agreement No. VPP-IZM-2018/2-0002 (subproject “Acquisition of Latvian Language”).

Keywords: corpus; learner corpus; target hypothesis; language acquisition; error annotation; corpus linguistics.