10.35539/LTNC.2022.0047.I.A.R.D.I.K.K.L.P.K.P.264.283

Valodas korpusu izmantošana latviešu valodas uzdevumu automātiskā ģenerēšanā

Ilze Auziņa, Roberts Darģis,
Inga Kaija, Kristīne Levāne-Petrova,
Kristīne Pokratniece

Šis darbs ir FLPP projekta "Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums" (projekta Nr. lzp-2018/1-0527) rezultāts sinerģijā ar Valsts pētījumu programmas "Humanitāro zinātņu digitālie resursi" projektu Nr. VPP-IZM-DH-2020/1-0001.

Atslēgvārdi: korpuslingvistika, valodas korpuss, latviešu valodas apguve, mācību uzdevumi, teikumu atlase

Ievads

Mūsdienās valodas korpusi nav tikai pētījumu empīrisks pamats, bet tie var būt noderīgi dažādu datos balstītu mācību materiālu un rīku izveidē. Korpusu izmantošanai valodas izpētē un apguvē, mācību un metodisko materiālu izstrādē ir lielas priekšrocības, jo korpusi piedāvā autentisku valodas materiālu un objektīvu skatījumu uz valodas parādībām, vairākkārt izmantojamus datus un iespēju pārbaudīt vai apstrīdēt hipotēzes par kādu valodas parādību vai valodu kopumā.

Lai pētītu valodu apguves īpatnības, analizētu valodas apguvēju pieļautās kļūdas un nodrošinātu datos balstītu valodas mācību un metodisko materiālu izstrādi, tiek veidoti valodas apguvēju korpusi, kuros tiek iekļauti valodas apguvēju (gan svešvalodas, gan otrās valodas) teksti un/vai runas dati (Leech 1998: xiv; Nesselhauf 2005: 40; Granger 2003). Valodas apguvēju korpusi ir specializēti korpusi, t. i., korpusi, kuros dati apkopoti pēc noteiktas pazīmes. Arī Latvijā pēdējo gadu laikā ir tapuši vairāki latviešu valodas apguvēju korpusi, t. sk. otrās baltu (latviešu vai lietuviešu) valodas apguvēju korpuss ESAM[1], "Valsts valodas prasmes pārbaudes darbu korpuss" (VVPP)[2] un jaunizveidotais "Latviešu valodas apguvēju korpuss" (turpmāk tekstā – LaVA)[3] (Auziņa u. c. 2021; Darģis et al. 2020), kura datu analīze ir pamatā rakstā aprakstītās pašmācības uzdevumu kopas izstrādē. Tā ir pirmā latviešu valodas apguvei paredzētā pašpārbaudes uzdevumu kopa, kurā iekļautie uzdevumi tiek ģenerēti automātiski no vairākiem valodas korpusiem un atbilst latviešu valodas prasmes pamatlīmenim.

Ar dabiskās valodas apstrādes (NLP) rīkiem (konformanču programmu, morfoloģisko analizatoru, tekstu sastatīšanu u. c.) var daļēji vai pilnībā automatizēt vairākus procesus, kas saistīti ar svešvalodu un otrās valodas (L2) apguvi. Viens no procesiem, ko arvien biežāk veiksmīgi automatizē, ir pašmācības uzdevumu ģenerēšana (O'Keeffe 2007; Pilán et al. 2016; Pilán et al. 2017; Smith et al. 2010; Volodina et al. 2012 u. c.). Uzdevumu izveidē tiek izmantoti ne tikai valodas apguvēju korpusi, bet arī paralēlie korpusi, kuros savstarpēji sastatīti teikumi un to tulkojumi citā valodā (Zanetti 2020). Par pamatu ņemot valodas korpusos atrodamos teikumus, leksiku, analizējot valodas apguvēju pieļautās kļūdas, tiek automātiski ģenerēti uzdevumi gan pašvadītai valodas apguvei, gan izmantošanai mācību procesā.

1 Otrās baltu valodas apguvēju korpuss ESAM http://esam.korpuss.lv

2 Valsts valodas prasmes pārbaudes darbu korpuss http://www.korpuss.lv/id/VVPP

3 Latviešu valodas apguvēju korpuss (LaVA) http://lava.korpuss.lv

Ja ir izstrādāta atbilstoša datu atlases metodika un kritēriji, izmantojamo korpusu skaits var būt neierobežots. Tādējādi tiek nodrošināta uzdevumu daudzveidība un atbilstība konkrētam mērķim, turklāt ir iespējams izvairīties no subjektivitātes.

Lai varētu automātiski ģenerēt uzdevumus, ir jāveic pamatīga datu analīze un jānosaka uzdevumos iekļaujamo vārdu, vārdformu un teikumu atlases kritēriji un atbilstība noteikta valodas prasmes līmeņa vingrinājumu elementiem. Izvēloties teikumus no korpusiem, ir vairāki papildu aspekti, kas jāņem vērā – 1) vai teikums ir saprotams izolēti, ārpus plašāka konteksta, 2) vai teikuma struktūra un lingvistiskā sarežģītība ir piemērota atbilstošam valodas prasmes līmenim. Arī runājot par tādu uzdevumu automātisku izveidi, kuros tiek piedāvāta lokāmo vārdšķiru vārdu paradigmu apguve, jāņem vērā lingvistiskā atbilstība atbilstošam valodas prasmes līmenim.

Uzdevumu izveidē ir vairāki posmi: 1) kvantitatīva un kvalitatīva LaVA kļūdu analīze un tipisko kļūdu noteikšana, 2) lingvistisko paraugteikumu ieguve no dažādiem latviešu valodas korpusiem, t. sk. "Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa" (LVK2018) un "Skolēnu pārspriedumu korpusa" (SPK), ietverot tikai LaVA korpusā lietotos vārdus, 3) dažādu tipu uzdevumu ģenerēšana, izmantojot atlasītos paraugteikumus un vārdu sarakstus.

Datu analīzē un atlasē izmantotas korpuslingvistikas metodes. Kvantitatīvai LaVA datu analīzei izgūti dažādi statistiskie radītāji, t. sk. absolūtais un relatīvais vārdu un vārdformu biežums korpusā, vārdu rangs. Kvalitatīvai datu analīzei un arī atbilstošu teikumu atlasei izmantotas konkordances.

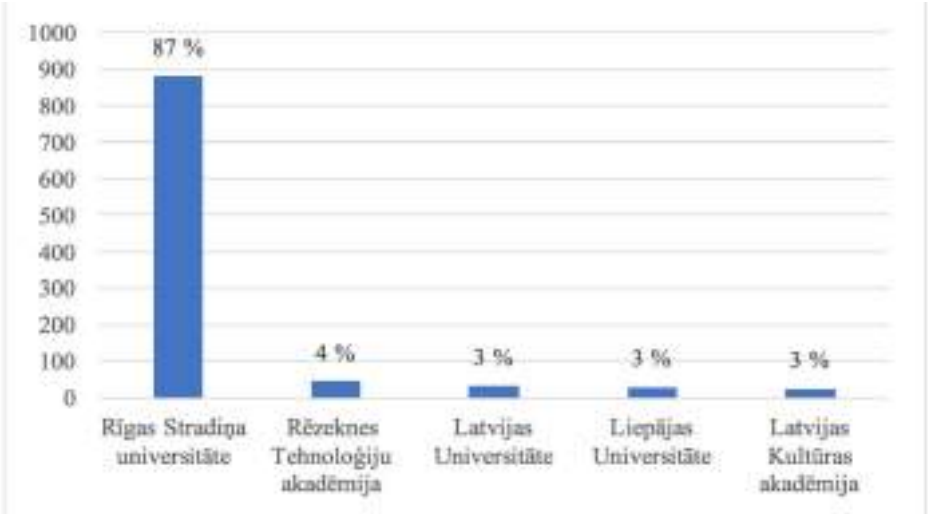
Raksta 1. nodaļā tiek raksturots "Latviešu valodas apguvēju korpuss" un analizēti tā dati, raksta 2. nodaļā aprakstīta uzdevumos iekļaujamo un izmantojamo valodas vienību atlases metodoloģija, savukārt raksta 3. nodaļā – uzdevumu kopas izstrāde un tajā iekļauto uzdevumu veidi.

LaVA korpuss

LaVA korpusa vispārīgs raksturojums

LaVA korpusā, kas fundamentālo un lietišķo pētījuma projektā (FLPP) "Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums" (projekta Nr. lzp-2018/1-0527)[4] laikā (2018–2021) izstrādāts LU MII, ir iekļauti 1015 to Latvijas augstākajās mācību iestādēs studējošo ārvalstnieku darbi, kuri latviešu valodu apgūst kā svešvalodu pirmo vai otro semestri, sasniedzot A1 (iespējams, A2) latviešu valodas prasmes līmeni. Korpusa LaVA apjoms ir gandrīz 190 000 vārdlietojumu.

4 Finansējuma avots: FLPP projekts "Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums" (Nr. lzp-2018/1-0527).



1. attēls. Datu ieguves avoti.

LaVA dati ir iegūti no piecām Latvijas augstskolām: Rīgas Stradiņa universitātes (RSU), Latvijas Universitātes (LU), Latvijas Kultūras akadēmijas (LKA), Liepājas Universitātes (LiepU) un Rēzeknes tehnoloģiju akadēmijas (RTA). Visvairāk tekstu ir no RSU (87 %, 882 teksti), no pārējām augstskolām daudz mazāk: RTA – 45 teksti (4 %), LU – 33 teksti (3 %), LiepU – 30 teksti (3 %) un LKA – 25 teksti (3 %), sk. 1. att.

Korpusu LaVA veido datu kopa (sk. http://lava.korpuss.lv/corpus/), kurā ir 1) studentu rakstītās esejas kopija, 2) digitalizētā eseja, 3) labotais teksts. Turklāt katram tekstam ir pievienota informācija jeb metadati par personas dzimumu, vecumu, dzimto valodu, laiku, cik ilgi tiek apgūta latviešu valoda (sk. 2. att.).

Tā kā lielākā daļa tekstu ir rokrakstā, tie vispirms tika digitalizēti, manuāli pārrakstot datorrakstā. Detalizētāk par korpusa LaVA izveides procesu sk. Levāne-Petrova et al. 2020. Savukārt digitalizētie teksti pēc korpusa veidotāju izstrādātās metodikas ir pārrakstīti atbilstoši valodas normām, t. i., ir izvirzīta mērķhipotēze (Auziņa u. c. 2020). Tādējādi teksti ir laboti, interpretējot valodas apguvēju rakstīto un izlemjot, ko valodas apguvējs katrā konkrētā gadījumā ir gribējis teikt, piem., *Es vinmer ēdu rīsu.* → *Es vienmēr ēdu rīsus.*

Gan oriģinālie, gan labotie teksti ir automātiski morfoloģiski marķēti – katrai vārdformai pievienojot morfoloģiskās pazīmes un pamatformu, piem., vārdformai *Rīgā* tiek pievienotas morfoloģiskās pazīmes [npfsl4], kur *n* ir lietvārds, *p* – īpašvārds, *f* – sieviešu dzimte, *s* – vienskaitlis, *l* – lokatīvs, *4* – 4. deklinācija, un norādīta pamatforma – *Rīga.* Automātiskā

2. attēls. Korpusa LaVA saskarnes ekrānuzņēmums.

3. attēls. Korpusa LaVA anotēšanas saskarne.

morfoloģiskā marķēšana ir veikta, izmantojot latviešu valodai izstrādāto automātiskās morfoloģiskās anotēšanas rīku (Paikens et al. 2013) un latviešu valodas morfoloģisko pazīmju kopu[5]. Tā kā latviešu valoda ir morfoloģiski daudznozīmīga un automātiskās marķēšanas procesā reizēm nav iespējams atšķirt konkrētas vārdformas vārdšķiru, dažkārt novērojamas kļūdas morfoloģiskajā marķējumā, un morfoloģiskais marķējums tika manuāli pārskatīts. Lai marķētu valodas apguvēju oriģinālajos tekstos visas lietotās tekstvienības, piem., paradigmai neatbilstošas formas (*universitātī, Franciji*) vai neeksistējošus vārdus un vārdformas (*bizeni, dārz*), ir izmantota reducēta morfoloģisko pazīmju kopa, parasti norādot tikai vārdšķiras identifikatoru: *Franciji* – tag [n], lemma [Francija]; *universitātī* – tag [n], lemma [universitāte]; *aizguļošamies* – tag [v], lemma [aizgulēties].

5 Korpuss.lv latviešu valodas morfoloģisko pazīmju kopa (v.2.2.1 (22.09.2021.)) http://www.korpuss.lv/LV_TagSet_v.2.2.1_22092021.pdf

Korpusā LaVA ir arī marķētas valodas apguvēju pieļautās kļūdas atbilstoši noteiktiem kļūdu tipiem: pareizrakstības kļūdas (*Spelling*), formveidošanas un vārddarināšanas kļūdas (*WordFormation*), leksikas kļūdas (*Lexical*), interpunkcijas kļūdas (*Punctuation*), sintakses kļūdas (*Syntactic*), kombinētas kļūdas (*Complex*). Sastatot korpusa oriģinālos un labotos tekstus, daļa kļūdu – pareizrakstības, formveidošanas un vārddarināšanas, interpunkcijas un leksikas kļūdas – ir marķētas pusautomātiski, pēc tam marķējumu manuāli pārskatot (sk. 3. att.). Savukārt sintakses un kombinētās kļūdas ir marķētas manuāli.

LaVA korpusā lietoto vārdu un vārdformu lietojuma biežums

Korpusa LaVA iekļautajos 1015 tekstos – gan oriģinālajos, gan labotajos – ir sastopami vairāk nekā 192 tūkstoši tekstvienību (t. sk. vārdformu, pieturzīmju, simbolu). Valodas apguvēji lietojuši 8400 dažādu vārdu un vairāk nekā 20 tūkstošus dažādu vārdformu.

Biežāk lietotie 20 vārdi korpusā ir *es* (16198), *būt* (11247), *un* (8220), *mana* (4119), *viņš* (3996), *patikt* (3886), *mans* (3580), *viņa* (2874), *dzīvot* (2476), *gads* (2365), *Rīga* (2280), *ar* (2071), *universitāte* (2044), *studēt* (1895), *saukt* (1866), *mēs* (1844), *uz* (1585), *bet* (1581), *no* (1538), *ļoti* (1532). Biežāk lietotie vārdi ir cieši saistīti ar studentu eseju tematiem – "Es un mana ģimene" (piem., *ģimene, brālis, draugs, māsa, tēvs, māte; saukt, studēt; liels, vecs, mīļš, mazs*), "Mana ikdiena" (piem., *studēt, iet, braukt; brīvs*), "Manas studijas" u. c.

	Lietvārds	Darbības vārds	Īpašības vārds	Vietniekvārds	Skaitļa vārds
1.	gads	būt	labs	es	divi
2.	Rīga	patikt	liels	mana	viens
3.	universitāte	dzīvot	vecs	viņš	divdesmit
4.	ģimene	studēt	mīļš	mans	trīs
5.	brālis	saukt	brīvs	viņa	astoņi
6.	draugs	ēst	jauns	mēs	četri
7.	māsa	runāt	mazs	tas	pieci
8.	māja	iet	skaists	katra	septiņi
9.	tēvs	garšot	grūts	tā	deviņi
10.	māte	braukt	nākamais	viss	seši

1. tabula. Biežāk lietoto lokāmo vārdšķiru vārdi.

Likumsakarīgi, ka vērojama korelācija starp vārdu lietojuma biežumu un kļūdaini lietotajām vārda formām, piem., visbiežāk valodas apguvēji kļūdījušies, lietojot lietvārdu *māja, brālis, darbs, draudzene, universitāte, māsa, nodarbība, bioķīmija, grāmata, bibliotēka* formas, no kuriem vārdi *gads, universitāte, brālis, māsa, māja* ir pirmajā desmitniekā.

LaVA korpusa kļūdu analīze

Izmantojot kļūdu marķējumu, *SketchEngine* programmā (Rychlý et al. 2007) ir iespējams iegūt informāciju par valodas apguvēju pieļautajām kļūdām. Kļūdu analīze liecina, ka apguvēju tekstos izplatītākās ir formveidošanas un vārddarināšanas kļūdas (40 %) un pareizrakstības kļūdas (33 %) (sk. 4 att.).

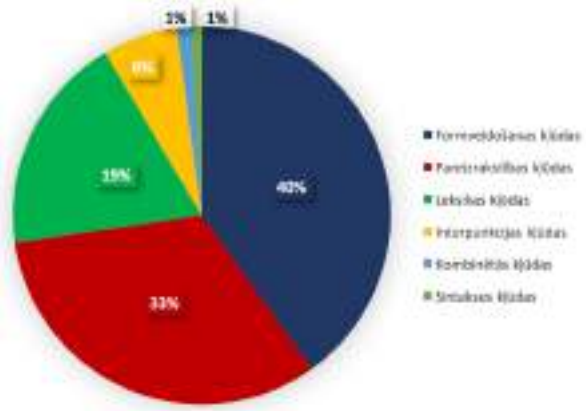
Kļūdaini lietoto vārdformu biežums korelē ar vārdu lietojumu biežumu korpusā, t. i., formveidošanas un vārddarināšanas kļūdas visvairāk sastopamas LaVA korpusā biežāk lietotajos vārdos: *mans, mana, viņš, gads, es, viņa, brālis, draugs, valoda, māte, māsa, divi, tēvs, Rīga, ģimene.* Formveidošanas un vārddarināšanas kļūdu grupā 61,7 % gadījumu ir kļūdaini lietoti lietvārdi, 17,7 % gadījumu – vietniekvārdi, 9,7 % gadījumu – darbības vārdi, 4,2 % gadījumu – īpašības vārdi, 3,9 % gadījumu – skaitļa vārdi un 2,8 % gadījumu – citu vārdšķiru vārdi. Visbiežāk valodas apguvēji nav atbilstoši lietojuši vienskaitļa akuzatīva formas 4. deklinācijas (*medicīnu, māsu, valodu, dienu, zobārstniecību, kafiju, anatomiju, bibliotēku* utt.) un 1. deklinācijas lietvārdiem (*tēvu, draugu, futbolu, autobusu, gadu, sieru* utt.). Nav pareizi lietotas arī 4. deklinācijas lietvārda vienskaitļa lokatīva, 1. konjugācijas vienskaitļa un daudzskaitļa nominatīva formas.

Visbiežāk valodas apguvēji kļūdās vārdu *patikt* (*patik → patīk, patick → patīk, patīka → patika* u. c.), *dzīvot* (*dzivo → dzīvo, dzivot → dzīvot* u. c.), *viņš* (*vina → viņa, vinam → viņam, vini → viņi* u. c.), *studēt* (*stude → studē, studejam → studējam* u. c.), *Rīga* (*riga → Rīga, Riga → Rīga* u. c.), *ļoti* (*loti → ļoti, lotī → ļoti* u. c.), *arī* (*ari → arī*), *universitāte* (*universitate → universitāte, universitatē → universitātē* u. c.), *viņa* (*vina → viņa, Vina → Viņa* u. c.), *ēst* (*ed → ēd, edīšu → ēdīšu, edu → ēdu* u. c.), *mēst* (*mes → mēs, musu → mūsu* u. c.), *runāt* (*runa → runā, runajam → runājam* u. c.), *nepatikt* (*nepatik → nepatīk* u. c.), *ģimene* (*giemene → ģimene, gimene → ģimene, gimenē → ģimenē* u. c.) formu pareizrakstībā – visbiežāk tās ir ar diakritisko zīmju lietojumu (neesamību) saistītas kļūdas.

Lielākā daļa pareizrakstības, formveidošanas un vārddarināšanas, kā arī leksikas kļūdu ir saistītas ar vienu tekstvienību – parasti vienu vārdu vai vārdformu, bet ir analītiskas gramatiskās formas, kurās gramatiskās nozīmes izteikšanai tiek izmantots patstāvīgs vārds kopā ar kādu palīgvārdu vai palīgvārda funkcijā lietotu vārdu, piem., darbības vārdu salikto laiku formas (*esmu lasījis, ir bijis* u. tml.), prepozicionāli savienojumi (*ar draugiem, uz skolu* u. tml.). Kļūdas analītisko gramatisko formu lietojumā marķētas kā sintakses kļūdas.

Latviešu valodā atšķirībā, piem., no angļu valodas, pieturzīmju lietojums teikumā ir ļoti nozīmīgs. Latviešu valodā interpunkcija ir balstīta uz gramatikas principiem, un atšķirīga pieturzīmju izmantošana bieži pilnībā maina teikuma nozīmi. Tādēļ korpusā, lai gan iesācēju līmenī interpunkcijas zināšanas valodas apguvējiem netiek prasītas, labotajos tekstos pieturzīmes liktas atbilstoši latviešu valodas interpunkcijas likumiem. Tas savukārt ļauj automātiski noteikt interpunkcijas kļūdas.

Bieži vienā vārdformā ir sastopamas vairāku tipu kļūdas – pareizrakstības un formveidošanas kļūda.



4. attēls. Kļūdu izplatība korpusā LaVA.

Paraugteikumu atlase un datu sagatavošana uzdevumu ģenerēšanai

Balstoties uz korpusa LaVA datu, t. sk. valodas apguvēju kļūdu, analīzi, pēc noteiktas metodoloģijas tiek ģenerēti uzdevumi. Uzdevumi paredzēti, lai palīdzētu valodas apguvējam nostiprināt latviešu valodas lingvistisko kompetenci, piem., darbības vārdu personu formu lietojumu īstenības izteiksmes vienkāršajos un saliktajos laikos, vārdu saskaņošanu, prievārdisko konstrukciju lietojumu.

Lai mācību uzdevumus varētu ģenerēt automātiski, tiek apzināts, kādas lingvistiskās prasmes valodas apguvējiem pamatlīmenī (A1, A2) jāsasniedz (sk. 2.1). Tas savukārt, izmantojot morfoloģiskās pazīmes, ļauj atlasīt teikumus ar atbilstošām vārdformām (sk. 2.3.) un LaVA korpusā sastopamajiem vārdiem (sk. 2.2.).

Lingvistisko prasmju definēšana

Šobrīd LaVA korpusā apkopotie dati atspoguļo to valodas apguvēju latviešu valodas prasmi, kuri tikai sāk apgūt valodu un iegūst pamatzināšanas, mācību laikā sasniedzot pamatlīmeņa 1. pakāpi (*Breakthrough*, (A1)) vai 2. pakāpi (*Waystage,* (A2)). Līdz ar to arī korpusā balstītie pašpārbaudes uzdevumi ir paredzēti tiem, kas latviešu valodu apgūst A1–A2 līmenī.

A1 ir zemākais valodas prasmes līmenis, ar kuru sākas jaunas, šajā gadījumā – latviešu, valodas apguve. A1 līmenī valodas apguvējs ir apguvis vienkāršas valodas formas (piem., darbības

vārda vienkāršās tagadnes, pagātnes un nākotnes formas, lietvārdu locīšanas paradigmu), spēj jautāt un atbildēt uz vienkāršiem jautājumiem par sevi, dzīvesvietu, pazīstamiem cilvēkiem u. tml., spēj veidot vienkāršus izteikumus par labi zināmiem tematiem, lieto ļoti vienkāršus valodas līdzekļus, nodarbībās apgūtu un ar konkrētiem tematiem un situācijām saistītu leksiku un frāzes. Valodas sniegums vēl ir ļoti nepilnīgs (Šalme, Auziņa 2016: 17). Savukārt A2 līmenī mācās izmantot valodu sociālo funkciju un komunikatīvās darbības īstenošanai, turpina pilnveidot valodas prasmes, mācoties saziņā lietot vienkāršas pieklājības formas, apjautāties par saziņas partnera sajūtām un emocionālo stāvokli, uzzināt jaunumus un iegūt citu informāciju, uzdot jautājumus un atbildēt uz jautājumiem par dažādām tēmām (Šalme, Auziņa 2016: 17).

Korpusā LaVA iekļauto tekstu atšķirības daļēji nosaka tas, ka tie tapuši dažādās augstskolās vai vienas augstskolas dažādu studiju kursu laikā, piem., no Rīgas Stradiņa universitātes tekstus ievākuši kursu "Latviešu valoda medicīnā", "Latviešu valoda zobārstniecībā" un "Latviešu valoda apmaiņas studentiem" docētāji. Katra kursa saturs var būt atšķirīgs atkarībā no kursa apjoma un apguvēju profesionālajām vajadzībām (Laizāne 2018: 36). Arī tekstu rakstīšanas laiks, t. i., 1. vai 2. semestris, semestra sākums vai beigas, ietekmē tekstos iekļauto saturu, it sevišķi gramatikas aspektā.

Tekstos atrodamās gramatiskās formas un konstrukcijas lielākoties atspoguļo līdz rakstīšanas brīdim kursā apgūto. Tā kā dažādos kursos un dažādos viena kursa posmos tas var būtiski atšķirties, uzdevumu veidošanai tika izveidots iekļaujamo gramatisko vienību saraksts, proti, tas veidots pēc A1–A2 līmeņa apraksta (Šalme, Auziņa 2016). Saskaņā ar šo sarakstu uzdevumos var būt iekļauti:

- 1.–6. deklinācijas lietvārdi;
- darbības vārdi darāmās kārtas īstenības izteiksmē;
- īpašības vārdi ar nenoteiktajām galotnēm;
- vietniekvārdi:
  - personu vietniekvārdi nominatīvā, ģenitīvā (izņemot *manis* un *tevis*), datīvā un akuzatīvā;
  - norādāmie, piederības, jautājamie, nenoteiktie un noteiktie vietniekvārdi nominatīvā, ģenitīvā, datīvā, akuzatīvā un lokatīvā;
  - attieksmes, atgriezeniskais vietniekvārds un nolieguma vietniekvārdi netiek iekļauti;
- apstākļa vārdi;
- prievārdi prepozitīvā lietojumā ar patstāvīgajiem vārdiem vienskaitlī:
  - *aiz, virs, zem, pie, no, ārpus, pirms, pēc, kopš, bez* saistījumā ar vienskaitļa ģenitīva formām;
  - *līdz* ar vienskaitļa datīva formām;
  - *ap, gar, pa, caur, pret, starp, pār, ar, par* ar vienskaitļa akuzatīva formām;

    - *uz* ar vienskaitļa akuzatīva vai vienskaitļa ģenitīva formām;
- vienkārši saikļi *un, bet, vai, ka, jo, tāpēc ka, ja, kā, lai*;
- pamata vai kārtas skaitļa vārdi (vienkārši vai salikteņi) visos locījumos, izņemot vokatīvu;
- izsauksmes vārdi;
- vienkāršas partikulas *vai* (tikai kā jautājuma partikula), *jā, nē, varbūt, arī, diemžēl, kā* (tikai salīdzinājumā), *nekā*
- saīsinājumi netiek iekļauti.

Korpusā LaVA iekļauto tekstu tematiku nosaka docētāja uzdevums (temats, par kādu autoriem uzdots tekstu rakstīt), piem., "Es un mana ģimene"; "Manas mājas un mana ikdiena"; "Mani vaļasprieki" u. c. Šie temati atbilst A1–A2 līmenim, taču tas, kādus konkrētus valodas elementus izvēlēties, ir autora ziņā. Tā kā, tekstus rakstot, ir atļauts izmantot palīglīdzekļus, piem., vārdnīcas, tekstos iekļautā leksika neaprobežojas ar A1–A2 līmenim tipisko, bet gan iekļauj tādu leksiku, kāda katram autoram ir šķitusi nepieciešama temata izklāstā. Tas nozīmē, ka pēc šī korpusa nevar izdarīt visaptverošus secinājumus par to, kādu leksiku autori zina, bet var pētīt, kādu leksiku apguvēji uzskata par lietderīgu un tiecas izmantot, rakstot par noteiktiem tematiem. Tāpēc leksikas ziņā uzdevumi ir balstīti nevis A1–A2 līmeņa aprakstā, bet gan korpusa LaVA datos.

Vārdu saraksta izveide

Uzdevumu ģenerēšanā ir svarīgi izmantot tikai tos vārdus, kuri valodas apguvējam būtu jāzina. Visprecīzāko informāciju par vārdiem, kurus apguvēji izmanto, var iegūt no apguvēju korpusa.

No apguvēju korpusa tika apkopota informācija par lemmām jeb pamatformām, kuras apguvēji ir izmantojuši tekstos. Apguvēju tekstos var parādīties arī specifiski vārdi, kuri netiek mācīti iesācēju līmenī, bet ko kāds apguvējs ir atradis vārdnīcā, jo tas bijis nepieciešams viņa tekstā. Lai izvairītos no šādu vārdu iekļaušanas vārdu sarakstā, tajā tiek iekļauti tikai tie vārdi, kuri korpusā sastopami vismaz trīs darbos.

Apguvēju korpusa apjoms nav pietiekams, lai iegūtu reprezentatīvu statistiku par katra vārda izmantotajām vārdformām, tāpēc iegūtie vārdi tika automātiski locīti un iegūtās vārdformas tika filtrētas, balstoties uz līmeņa aprakstu. Iegūtais vārdformu saraksts tika tālāk izmantots uzdevumu ģenerēšanā un paraugteikumu atlasē.

Teikumu atlase

Daļai uzdevumu tipu ir nepieciešami teikumi. Tā kā korpusā LaVA atrodamie teikumi ir ar dažādiem ierobežojumiem (sk. rakstā iepriekš), lai nodrošinātu pietiekami lielu teikumu

dažādību, teikumi tika atlasīti arī no korpusa LVK2018 un SPK. Teikumu atlasei izmantotais tekstu korpuss LVK2018 ir aptuveni 10 miljonu vārdlietojumu liels, vispārīgs korpuss, kas sastāv no dažādu žanru tekstiem (periodikas (60 %), daiļliteratūras (20 %), zinātniskiem tekstiem (10 %), normatīvajiem aktiem (8 %) un Saeimas stenogrammām (2 %)). LVK2018 ir automātiski morfoloģiski marķēts, un tam pievienoti metadati (Levāne-Petrova 2012; 2019). Savukārt SPK ir veidots, lai pētītu skolēnu latviešu valodas prasmes un nodrošinātu datos balstītu latviešu valodas mācību un metodisko materiālu izstrādi. SPK ir iekļauti 468 pārspriedumi no 12. klases latviešu valodas eksāmenu darbiem (apmēram 185 000 vārdlietojumu). Tie ir latviešu mācībvalodas vidusskolu, mazākumtautību skolu un valsts ģimnāziju audzēkņu darbi no Kurzemes, Latgales un Rīgas. SPK ir iekļauti nelaboti teksti, paturot visas skolēnu pieļautās drukas, interpunkcijas u. c. kļūdas (Levāne-Petrova, Pokratniece 2021).

Atlasītajiem teikumiem jāatbilst valodas apguvēju līmenim. Lai nodrošinātu, ka teikumos izmantotā leksika ir saprotama valodas apguvējiem, pirmajā kārtā tika atlasīti tikai tie teikumi, kuros bija vārdi un vārdformas tikai no iepriekš izveidotā saraksta.

Ar teikuma filtrēšanu tikai pēc vārdu saraksta bieži vien nepietiek. Starp atlasītajiem teikumiem ir gan pārāk gari un sarežģīti teikumi, gan arī pārāk īsi teikumi, kas nesatur pietiekami daudz informācijas uzdevumu ģenerēšanai, piem., teikumi ar nepilnu struktūru. Nākamajā teikumu atlases kārtā ir nepieciešams veikt sintaktisko filtrēšanu, lai izvairītos no šādu teikumu iekļaušanas uzdevumu ģenerēšanā. Filtrēšana tika veikta, balstoties uz šādiem kritērijiem:

- viens teikums nevar sastāvēt vairāk nekā no trim neatkarīgām teikuma daļām;
- visās teikuma daļās jābūt teikuma gramatiskajam centram – vēlams, gan teikuma priekšmetam, gan izteicējam;
- neiekļaut vienā teikuma daļā vairāk par diviem apstākļa vārdiem (tomēr tajā var būt vairāk nekā divi paplašinātāji);
- neiekļaut vienā teikumā vairāk par divām partikulām.

Kritēriji ir izveidoti atbilstoši docētāju pedagoģiskajai pieredzei, mācot latviešu valodu kā svešvalodu augstākās izglītības iestādēs.

Lietojot latviešu valodu kā dzimto valodu, teikumi mēdz būt aprauti, ar mainītu vārdu secību vai izlaistu informāciju, kas nosakāma no konteksta. Tāpēc no iepriekš minētajos valodas korpusos iegūtajiem teikumiem tika atlasīti 1000 teikumi (800 teikumi no LVK2018 un 200 teikumi no SPK), pēc nepieciešamības tos pielāgojot A1–A2 līmenim. Veiktas šādas izmaiņas:

- samazināta teikuma atkarība no konteksta (piem., *Un vai tieši Latvijas iedzīvotājiem vienmēr būs šī nauda? → Vai Latvijas iedzīvotājiem vienmēr būs nauda?*);
- pievienots teikuma priekšmets un/vai izteicējs (piem., *Gribu pastāstīt, kā šis jautājums virzās. → Es gribu pastāstīt, kā šis jautājums virzās.*);

- teikuma vārdu secība mainīta uz neitrālu (piem., *Nu, nebūs tā nekāda izvēles iespēja!* → *Tā nebūs nekāda izvēles iespēja*.);
- izmantota tikai daļa(-as) no sākotnējā teikuma (piem., *Tad, kad mēs teicām: "Bet mēs esam no Latvijas!"* → *Mēs esam no Latvijas*.);
- dažkārt teikums izmantots tikai ierosmei, veidojot tā vietā citu (piem., *Viņa, māte, to nebija varējusi*. → *Viņa māte to nevarēja*.)

Teikumus manuāli atlasīja un pēc vajadzības pārveidoja praktizējoša latviešu valodas kā svešvalodas docētāja pieaugušajiem augstākās izglītības iestādē Latvijā ar pieredzi mācību materiālu gatavošanā.

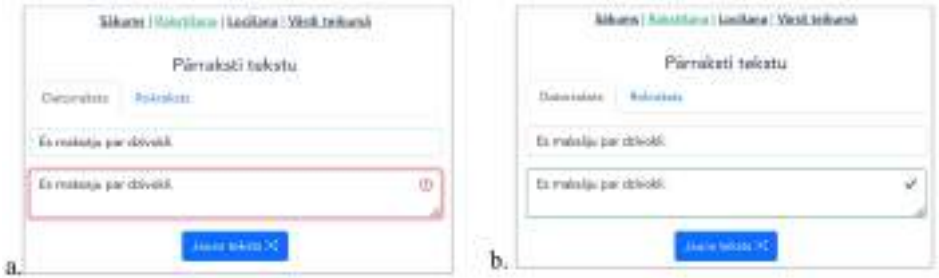
Uzdevumu tipi un uzdevumu ģenerēšana

Balstoties uz kļūdu analīzi, tika izvēlēti trīs dažādu uzdevumu veidi: teksta pārrakstīšanas uzdevumi (sk. 3.1.), vārdu locīšanas uzdevumi (sk. 3.2.) un uzdevumi, kuros valodas apguvējam teikumā jāieraksta vai jāizvēlas atbilstoša vārda forma (sk. 3.3.). Izraudzītie uzdevumu veidi atbilst biežāk sastopamajām apguvēju kļūdām un palīdz izvairīties no tām: rakstīšanas uzdevumi palīdz novērst pareizrakstības kļūdas, locīšanas uzdevumi – formveidošanas kļūdas, savukārt, ierakstot vai ievietojot atbilstošo vārdformu teikumā, valodas apguvēji nostiprina savas zināšanas formveidošanā un sintaksē. Leksikas un interpunkcijas apguvei uzdevumi nav veidoti.

Pārrakstīšanas uzdevumi (Rakstīšana)

Latviešu valodas rakstības pamatā ir latīņu alfabēts, kas papildināts ar diakritiskām zīmēm patskaņu garuma (¯), palatālo līdzskaņu (,), šņāceņu (ˇ) apzīmēšanai. Tieši šo papildu diakritisko zīmju kļūdaina izmantošana vai neesamība ir biežs pareizrakstības kļūdu cēlonis. Gan patskaņa garumam neatbilstoša burta izmantojums (īsa patskaņa vietā tiek rakstīts gara patskaņa burts, gara patskaņa vietā – īsa patskaņa burts), gan līdzskaņu rakstība bez diakritiskajām zīmēm vai neatbilstošas diakritiskās zīmes izmantojums, bieži vien ir saistīts ar valodas apguvēja dzimtās valodas fonētiski fonoloģiskās sistēmas un/vai grafētikas atšķirību no latviešu valodas (Auziņa u. c. 2019). Nepietiekamas zināšanas par latviešu valodas diakritiskajām zīmēm dažkārt noved arī pie latviešu valodas sistēmā neesošas diakritiskās zīmes lietojuma (Kaija 2020). Pareizrakstības kļūdas mēdz būt arī, piem., burtu rakstīšana nepareizā secībā, it sevišķi garākos vārdos.

Teksta pārrakstīšana palīdz labāk apgūt latviešu valodas grafētisko sistēmu, gan vizuāli pievēršot uzmanību burtu secībai un diakritiskajām zīmēm, gan šo secību atkārtojot pašrocīgi. Arī apguvējiem ar disleksiju pārrakstīšana tiek minēta kā viens no noderīgiem pareizrakstības apguves veidiem (Crombie 2000). Šādi uzdevumi arī ļauj trenēt diakritisko zīmju lietošanu datorrakstā, novēršot gadījumus, kad diakritiskās zīmes netiek lietotas tikai tāpēc, ka apguvējs tehniski nav pieradis to darīt.



5. attēls. Rakstīšanas uzdevuma saskarne: kļūdaina izpilde (a) un pareiza izpilde (b).

Pārrakstīšanas uzdevumos valodas apguvējam ir iespēja izvēlēties, vai pārrakstīt jau datorrakstā dotu teikumu vai rokrakstā. Datorrakstā tiek piedāvāts nejauši izvēlēts teikums no iepriekš atlasītajiem (sk. 2.3.). Rokrakstā tiek piedāvāts nejauši izvēlēts teikuma attēls, kas iegūts manuāli, izgriezts no apguvēju darbiem.

Ja valodas apguvējs, pārrakstot teikumu, kļūdās, par to uzreiz tiek ziņots – rāmis ap tekstu kļūst sarkans (sk. 5. (a) att.). Par kļūdu tiek uzskatīts nepareizs burts, burts bez diakritiskās zīmes vai ar lieku/neatbilstošu diakritisko zīmi, neatbilstošs lielo un mazo burtu lietojums, lieka vai neesoša atstarpe, lieka, trūkstoša vai nepareiza pieturzīme. Kad viss teikums ir pārrakstīts pareizi, tas tiek iekrāsots zaļā krāsā. Jebkurā brīdī ir iespējams pieprasīt jaunu teikumu (sk. 5. (b) att.).

Vārdu locīšanas uzdevumi (Locīšana)

Otra grupa ir uzdevumi, kuros valodas apguvējam tiek piedāvāts apgūt latviešu valodas lokāmo vārdšķiru – lietvārdu, darbības vārdu, īpašības vārdu, vietniekvārdu un skaitļa vārdu – locīšanas paradigmas. Uzdevumi aptver tikai daļu to gramatikas kategoriju, kuras atbilst latviešu valodas prasmes līmenī dotajam lingvistiskās kompetences raksturojumam, konkrēti – noteiktajām gramatikas zināšanām (Šalme, Auziņa 2016). Valodas apguvējs, izmantojot šos uzdevumus, var apgūt:

- 1.–6. deklinācijas lietvārdu locīšanu piecos locījumos – nominatīvā, ģenitīvā, datīvā, akuzatīvā un lokatīvā – vienskaitlī un daudzskaitlī, arī daudzskaitlinieku un vienskaitlinieku locīšanu atbilstošajās formās;
- I–III konjugācijas un nekārtno tiešo un atgriezenisko darbības vārdu locīšanu

darāmās kārtas īstenības izteiksmes vienkāršās tagadnes, pagātnes un nākotnes formās;

- nenoteikto īpašības vārdu locīšanu sieviešu un vīriešu dzimtes formās, saskaņojot tos ar atbilstošo lietvārdu dzimtē, skaitlī un locījumā;
- atsevišķu vietniekvārdu grupu, konkrēti – personas vietniekvārdu, norādāmo vietniekvārdu, attieksmes (arī jautājamo) vietniekvārdu –, locīšanu;
- pamata un kārtas skaitļa vārdu locīšanu, t. sk. skaitļa vārdu savienojumus, piem., *divdesmit viens, četrdesmit pieci.*

Izvēloties, kuras vārdšķiras vārdu locīšanu apgūt, tiek piedāvāta papildu izvēle, piem., lietvārdiem – deklinācija, darbības vārdiem – konjugācija, nekārtnie darbības vārdi, tiešie vai atgriezeniskie darbības vārdi. (sk. 6. att.). Mācoties īpašības vārda locīšanas paradigmas, tiek piedāvāts izvēlēties, ar kuras deklinācijas lietvārdu to kopā locīt, saskaņot dzimtē, skaitlī, locījumā. Locīšanai tiek izmantots korpusā LaVA sastopamo vārdu krājums (sk. 2.1.).

Pēc tam, kad ir izraudzīta vārdšķira un atbilstošā kategorija, no LaVA vārdu saraksta automātiski tiek piedāvāts vārds locīšanai (piem., "Izloki *nodarbība*"), kurš pēc morfoloģiskajām pazīmēm atbilst vēlamajam. Piem., ja valodas apguvējs vēlas apgūt 1. deklinācijas lietvārdu locīšanu, tiek atlasīti vārdi, kuru vārdformas identifikators, 1. pazīme, ir [n], savukārt 6. pazīme – [1], t. i., [n....1]. No atlasītajiem vārdiem pēc nejaušības principa tiek piedāvāts viens.

Pildot uzdevumu, tiek piedāvātas vairākas iespējas: 1) pārbaudīt ierakstītās vārdformas ("Pārbaudīt"), 2) apskatīt pareizās atbildes ("Parādīt atbildes"), 3) izvēlēties citus vārdus ("Cits"). Jebkurā laikā var aplūkot pareizās atbildes un pēc vēlēšanās tās atkal paslēpt, nospiežot pogu "Parādīt atbildes". To var darīt arī tad, ja valodas apguvējs nezina, kā piedāvāto vārdu pareizi locīt. Pēc pārbaudes pogas nospiešanas pareizās vārdformas tiek dotas uz zaļa fona, nepareizās vārdformas vai neaizpildītās vietas – iezīmētas ar sārtu fonu.

Vārdformu ievietošanas/ierakstīšanas uzdevumi (Vārdi teikumos)

Valodas apguvē nozīmīgi ir tukšo vietu aizpildīšanas uzdevumi (*gap-fill exercise*), kas tiek izmantoti gan leksikas, gan gramatikas apguvē. Šāda tipa uzdevumi tiek izmantoti arī LaVA uzdevumu kopā, ļaujot valodas apguvējiem ierakstīt atbilstošu vārdformu vai izvēlēties atbilstošo no dotajām.

Korpusi piedāvā plašu, ērti un ātri iegūstamu piemēru klāstu, kas atspoguļo reālo valodu, un ir pierādīts, ka korpusa piemēru izmantošana pozitīvi ietekmē valodas apguvēju lingvistisko prasmju attīstību. Protams, valodas apguvējiem vismaz iesācēju līmenī varētu būt grūti šos teikumus saprast. Otra alternatīva ir mācību līdzekļu autoru mākslīgi veidoti, *neautentiski* teikumi, kas balstās uz pedagogu pieredzi un intuīciju. Veidojot korpusā balstītus uzdevumus, kuros vārds teikumā jāievieto vai jāieraksta atbilstoša formā, mēģinām rast viduscelu – izmantojot skolotāju intuīciju un zināšanas, latviešu valodas prasmes līmeņa aprakstu un

6. attēls. Gramatisko kategoriju izvēle locīšanas uzdevumos.

7. attēls. Uzdevums 6. deklinācijas lietvārdu formu apguvei.

korpusā balstītus pētījumus, uzdevumos tiek ietverti mērķtiecīgi atlasīti, daudzveidīgi teikumi, kas tiek piedāvāti pēc nejaušības principa.

Teikumos tiek atrasti vārdi, kas ir vārdformu sarakstā, un tiek ģenerēts iespējamo vārdformu saraksts (ievietošanai). Teikumos var būt dažādu vārdšķiru vārdu dažādas formas, bet ievietošanai tiek piedāvātas tās pašas vārdformas, kuru apguvi trenēt ir iespējams locīšanas uzdevumos, tātad 1.–6. deklinācijas lietvārdi utt. Līdzīgi kā locīšanas uzdevumos, arī vārdformu ievietošanas uzdevumos piemēri tiek ģenerēti tikai ar tām vārdšķirām, kuras ir atlasītas izvēlnē.

Uzdevumā pēc nejaušības principa tiek parādīti desmit teikumi, kuros katrā ir viena tukša vieta. Teikumā ne vienmēr pietiek informācijas, lai noteiktu, kāda forma tajā bijusi sākotnēji (piem., vai darbības vārds lietots tagadnes, pagātnes vai nākotnes formā), tāpēc pirms tukšās vietas ir norādīta ievietojamā vārda pamatforma un papildu lingvistiskā informācija (piem., darbības vārdam – laiks, skaitlis, persona). To izmantojot, apguvējam jāieraksta tukšajā vietā pareizā vārdforma. (Sk. 7. att.)

Uzdevumu var pildīt arī kā daudzizvēļu uzdevumu. Aiz tukšā lauka ir poga ar jautājuma zīmi. Nospiežot to, atveras saraksts ar piecām attiecīgajām vārdformām, no kurām viena ir pareizā.

Nospiežot uz izvēlētās formas, tā automātiski tiek ierakstīta tukšajā vietā. Jebkurā laikā ir iespējams arī aplūkot vai paslēpt pareizās atbildes, atkārtoti nospiežot pogu "Parādīt atbildes", vai ģenerēt jaunus piemērus.

Nobeigums

Dažādi valodas korpusi, bet jo īpaši apguvēju valodas korpusi, ir izmantojami mācību uzdevumu izveidē. Ja apguvēju valodas korpuss ir marķēts dažādos līmeņos, uzdevumus, izmantojot dažādus dabiskās apstrādes un analīzes rīkus, kā arī atbilstošu metodoloģiju, ir iespējams veidot automātiski. Valodas apguvēju korpuss parāda, uz kādu valodas elementu pārbaudi vajadzētu koncentrēties, savukārt lielākai piemēru dažādībai var izmantot citus mērķvalodas korpusus, atlasot tajos tekstvienības, kas atbilst valodas apguvēju korpusā konstatētajām.

No korpusa LaVA, LVK2018 un SPK paraugteikumu kopas, kas tika izveidota pēc īpašiem kritērijiem un metodoloģijas, tika automātiski veidoti dažādu veidu pašpārbaudes uzdevumi latviešu valodas apguvei A1 un A2 līmenī. Balstoties uz korpusa LaVA kļūdu analīzi, tika izveidoti trīs dažādu uzdevumu tipi: teksta pārrakstīšanas uzdevumi, vārdu locīšanas uzdevumi un vārdformu ievietošanas un ierakstīšanas uzdevumi. Visu veidu uzdevumus valodas apguvējs var veikt patstāvīgi jebkurā laikā sev ērtā vietā, pēc vajadzības aplūkojot arī pareizās atbildes. Izveidotie mācību uzdevumi ir pieejami ikvienam interesentam korpusa LaVA tīmekļvietnes sadaļā "Uzdevumi" (skat. http://lava.korpuss.lv). Uzdevumu kopu paredzēts aprobēt, iesaistot latviešu valodas apguvējus, kas sāk mācīties latviešu valodu. Tāpat uzdevumu kopu ir iespējams papildināt ar citiem uzdevumu tipiem.

Automātiski ģenerēti uzdevumi ne tikai ļauj docētājam izmantot jau gatavu materiālu bez nepieciešamības to veidot no jauna vai patērēt laiku atbilžu pārskatīšanai un labošanai, bet arī sniedz iespēju apguvējam šos uzdevumus pildīt neierobežotu skaitu reižu. Turklāt tas, ka ir iespējams izvēlēties, piem., lietvārda deklināciju, no kuras atlasīt piemērus, ļauj šos uzdevumus izmantot atkārtoti dažādos valodas apguves posmos, pēc vajadzības atlasot attiecīgi apgūstamās lingvistiskās kategorijas vai arī visas līdz šim apgūtās. Arī uzdevuma grūtības pakāpi var samērot ar pedagoģiskā procesa vajadzībām, piem., uzdevumu ar formu ievietošanu teikumā pildīt kā daudzizvēļu uzdevumu. Uzdevumi ir izmantojami ne tikai docētāja vadītā mācību procesā, bet arī pašmācībā, piem., apgūstot noteiktu vārdšķiru locīšanas paradigmas. Plašās lietojuma iespējas ar iespējām pielāgot uzdevumus konkrēta apguvēja vajadzībām ir nozīmīga automātiski ģenerētu uzdevumu priekšrocība.

Šobrīd uzdevumi tiek piedāvāti tikai latviešu valodas apguvei A1 un A2 līmenī, tomēr, izmantojot rakstā aprakstītos principus, ir iespējams modelēt uzdevumus arī augstāku latviešu valodas prasmes līmeņu apguvei.

Bibliogrāfija

Auziņa, Ilze, Kaija, Inga, Levāne-Petrova, Kristīne, Darģis, Roberts, Pokratniece, Kristīne (2021). Latviešu valodas apguvēju korpusa (LaVA) izmantošana pētniecībā un mācību uzdevumu izstrādē. *Latviešu valodas apguve. XIII Starptautiskais baltistu kongress: rakstu krājums.* Liepāja: LiePA, 142.–161. lpp.

Auziņa, Ilze, Kaija, Inga, Levāne-Petrova, Kristīne (2020). Mērķhipotēžu izvirzīšana latviešu valodas apguvēju korpusā. *Valoda: nozīme un forma,* 11. Rīga: LU Akadēmiskais apgāds, 7.–26. lpp.

Auziņa, Ilze, Levāne-Petrova, Kristīne, Darģis, Roberts (2019). Latviešu valodas apguvēju kļūdu analīze: pareizrakstības kļūdas. Smiltniece, Gunta, Lauze, Linda (atb. red.). *Vārds un tā pētīšanas aspekti*, Nr. 23(1/2). Liepāja: LiePA, 220.–227. lpp.

Crombie, Margaret A. (2000). Dyslexia and the learning of a foreign language in school: where are we going? *Dyslexia*, No. 6(2), pp. 112–123.

Granger, Sylviane (2003). International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL* Quarterly, No. 37(3), pp. 538–546.

Kaija, Inga (2020). Jaunu burtu veidošana ar diakritiskajām zīmēm latviešu valodas kā svešvalodas apguvēju tekstos. *Valodu apguve: problēmas un perspektīva*: zinātnisko rakstu krājums. Liepāja: LiePA, 102.–110. lpp.

Laizāne, Inga (2018). Latviešu valodas kā svešvalodas apguve Baltijas valstīs. *Valodu apguve: problēmas un perspektīva*: zinātnisko rakstu krājums. Liepāja: LiePA, 28.–39. lpp.

Leech, Geoffrey (1998). Preface. *Learner English on computer*. Granger, Sylviane (ed.). London: Addison Wesley Longman, xiv–xx.

Levāne-Petrova, Kristīne, Auziņa, Ilze, Pokratniece, Kristīne (2020). Latviešu valodas apguvēju korpusa datu ieguves un apstrādes metodoloģijas izstrāde. *Valodu apguve: problēmas un perspektīva*: zinātnisko rakstu krājums. Liepāja: LiePA, 299.–309. lpp.

Levāne-Petrova, Kristīne (2012). Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji *Baltistica. VIII. Priedas.* Vilnius: Vilniaus Universiteto leidykla, 89.–98. lpp.

Levāne-Petrova, Kristīne (2019). Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *Valoda: nozīme un forma,* 10. Rīga: LU Akadēmiskais apgāds, 131.–146. lpp.

Levāne-Petrova, Kristīne, Pokratniece, Kristīne (2021). Skolēnu pārspriedumu korpusa izveide. *XIII Starptautiskais baltistu kongress.* Liepāja: Liepājas Universitāte, 106.–118. lpp.

Nesselhauf, Nadja (2005). *Collocations in a learner corpus.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

O'Keeffe, Anne, McCarthy, Michael, Carter Ronald (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge University Press.

Pilán, Ildikó, Vajjala, Sowmya, Volodina, Elena (2016). A readable read: automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications*, No. 7(1), pp. 143–159.

Pilán, Ildikó, Volodina, Elena, Borin, Lars (2017). Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *arXiv preprint arXiv:1706.03530*

Paikens, Pēteris, Rituma, Laura, Pretkalniņa, Lauma (2013). Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA).*

Rychlỳ, Pavel (2007). Manatee/Bonito-A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 65– 70.

Smith, Simon, Avinesh, P.V.S, Kilgarriff, Adam (2010). Gap-fill tests for language learners: corpus-driven item generation. *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing,* pp. 1–6.

Šalme, Arvils, Auziņa, Ilze (2016). *Latviešu valodas prasmes līmeņi: pamatlīmenis A1, A2, vidējais līmenis B1, B2*. Rīga: LVA.

Volodina, Elena, Johansson, Richard, Johansson, Sofie (2012). Semi-automatic selection of best corpus examples for Swedish: initial algorithm evaluation. *Proceedings of the Workshop on NLP for CALL*, Vol. 80, pp. 59–70.

Zanetti, Arianna (2020). *NLP methods for the automatic*

generation of exercises for second language learning from parallel corpus data. Master Thesis. Gothenburg University. Available: https://gupea.ub.gu.se/bitstream/handle/2077/66583/gupea_2077_66583_1.pdf?sequence=1&isAllowed=y [accessed 18.03.2022.].

Summary

Use of the Language Corpora in Automatic Generation of Latvian Language Exercises

Ilze Auziņa, Roberts Darģis, Inga Kaija, Kristīne Levāne-Petrova, Kristīne Pokratniece

Keywords: computational linguistics, language corpora, Latvian language acquisition, sentence selection, exercises

Today, language corpora are not only the empirical basis of research but can also be used in developing a variety of data-driven teaching materials and tools. The experience of other countries shows that the development of self-assessment exercises for language learning can be partially or fully automated using language corpora and natural language processing (NLP) tools, thus providing both a variety of exercises and support for teachers in the implementation of the curriculum.

The Latvian Language Learners Corpus (LaVA) developed at the Institute of Mathematics and Computer Science, University of Latvia, includes more than 1000 texts created by foreign Latvian language learners studying at Latvian higher education institutions for the first or second semester reaching A1 (possibly A2) Latvian language proficiency level. The size of the corpus is more than 180 000 words. According to the LaVA data analysis, including learners error analysis, exercises and tests are generated. Data analysis allows us to identify problematic spelling, grammar, and vocabulary issues. The exercises are intended to help the language learner to strengthen the linguistic competence of Latvian language, for example, the use of verb forms in the indicative mood, both in indefinite and perfect tense forms.

The article discusses the methodology according to which, based on the statistical and quantitative analysis of the LaVA corpus data, sample sentences are selected from different corpora of Latvian language, for example, *The Balanced Corpus of Modern Latvian* (LVK2018), *The Corpus of Students' Essays* (SPK), as well describes the task-development algorithms and development of online self-assessment exercises site.