Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos

The Balanced Corpus of Modern Latvian, its role in grammar studies

Kristīne Levāne-Petrova

Latvijas Universitātes
Matemātikas un informātikas institūts
Raiņa bulvāris 29, Rīga, LV-1050
E-pasts: kristine.levane-petrova@lumii.lv

Mūsdienās gramatiskie pētījumi vairs nav iedomājami bez tekstu korpusiem, jo īpaši vispārīgā valodas korpusa, kas noteiktās proporcijās atspoguļo valodu visā tās kopumā. Tādējādi vispārīgais valodas korpuss mūsdienās ir uzskatāms par gramatiskās domas pamatu, jo tajā tiek veikti pētījumi visos valodas līmeņos. Tāpat arī korpusa izveide ir cieši saistīta ar gramatiskajiem pētījumiem, piem., morfoloģiskā marķējuma izstrāde korpusam.

Šajā rakstā aplūkota „Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa” 2018. gada versija, tā uzbūve, tekstu atlasas kritēriji un citi ar korpusa uzbūvi saistīti jautājumi. Pats valodas korpuss ir datorizētai analīzei pieejams apjomīgs tekstu kopums, kas izveidots atbilstoši noteiktiem kritērijiem.

Raksta ievadā ir izklāstīta „Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa” attīstības vēsture, tā vispārīgs raksturojums. Tālāk rakstā ir aplūkota korpusa uzbūve, tekstu atlasas kritēriji, kā arī apskatītas korpusa konkrētas sadaļas un ar to izveidi saistītie jautājumi un risinājumi. Tāpat rakstā ir īsi raksturoti korpusa metadati, morfoloģiskais marķējums un meklēšanas iespējas korpusā.

Atslēgvārdi: valodas korpuss; tekstu korpuss; vispārīgais korpuss; tekstu atlasas kritēriji; līdzsvarotība; metadati.

Ievads

Mūsu dienās gramatiskie pētījumi ir grūti iedomājami bez dažādiem tekstu korpusiem, jo īpaši – vispārīgā korpusa,¹ kas noteiktās proporcijās atspoguļo valodu visā tās kopumā (Aston, Burnard 1998, 5). Savukārt pats vispārīgais korpuss nav iedomājams bez morfoloģiskā marķējuma (Paikens 2007, 2016; Paikens, Rituma, Pretkalniņa 2013), proti, katram vārdlietojumam ir automātiski pievienotas tam atbilstošas morfoloģiskās pazīmes un pamatforma (lemma). Šāds marķējums ļauj ievērojami ātrāk un vieglāk tekstu korpusā atrast nepieciešamos datus, piem., visus 6. deklinācijas lietvārdus latviešu valodā. Tādējādi mūsdienās vispārīgais korpuss ir pamats valodas izpētei visos tās līmeņos. Arī pati korpusa izveide ir cieši saistīta ar gramatiskajiem pētījumiem, jo korpusa morfoloģiskais marķējums atspoguļo formalizētu skatījumu uz valodu, kas nedaudz atšķiras no

¹ Angļu valodā iepriekš izmantotais termins *general corpus* (vispārīgais korpuss) laika gaitā ir aizstāts ar terminiem *reference corpus*, arī līdzsvarotais korpuss (*balanced corpus*).

tradicionālajiem gramatikas aprakstiem. Piem., korpusa morfoloģiskais marķējums neiekļauj noliegtos vietniekvārdus kā atsevišķu vietniekvārdu grupu un instrumentāli kā atsevišķu locījumu.

Jau vairāk nekā desmit gadus latviešu valodas izpēte dažādos līmeņos ir cieši saistīta ar „Līdzsvaroto mūsdienu latviešu valodas tekstu korpusu” (LVK), kas tiek izmantots kā empīrisks pamats visdažādākajos latviešu valodas pētījumos un darbos – sākot no studentu zinātniski pētnieciskajiem darbiem, beidzot ar jaunākās „Latviešu valodas gramatikas” (Nītiņa, Grigorjevs 2013) izstrādi. Tāpat ir arī publicēti raksti par LVK izmantošanas iespējām pētniecībā (piem., Andronova, Andronovs 2011; Levāne-Petrova 2011). Šis raksts ir veltīts „Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa”² 2018. gada versijas izstrādei, korpusa tekstu atlases kritērijiem, metadatiem, morfoloģiskajam marķējumam, kā arī citiem ar korpusa uzbūvi saistītiem jautājumiem. Tāpat rakstā ir īsi aplūkotas meklēšanas iespējas korpusā.

Šajā rakstā ar tekstu korpusu tiek saprasts strukturēts tekstu un/vai transkribētas runas krājums, kas paredzēts lingvistiskai analīzei un valodas tehnoloģiju izstrādei (sk., piem., Bussman 2006, 226; Skujiņa 2007, 196; Collins English Dictionary 2018).

1. „Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa” izveides vēsture un tā vispārīgs raksturojums

Darbs pie LVK tika sākts 2005. gadā, kad pēc „Valsts valodas aģentūras” (VVA) pasūtījuma Latvijas Universitātes Matemātikas un informācijas institūtā (LU MII) tika izstrādāta „Latviešu valodas korpusa koncepcija” (Levāne-Petrova 2012). Kopš 2007. gada ar „Latviešu valodas aģentūras” (LVA) atbalstu tika izveidotas vairākas LVK versijas, līdz 2013. gadā „Līdzsvarotais mūsdienu latviešu valodas tekstu korpus” sasniedza apmēram 4,5 miljonus vārdlietojumu (LVK2013).

Jaunā „Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa” versija (LVK2018) tika veidota, paplašinot „Līdzsvaroto mūsdienu latviešu valodas tekstu korpusu” (LVK2013).³ Tādējādi jaunās korpusa versijas izveidē ir ievēroti gandrīz visi LVK2013 uzbūves principi (Levāne-Petrova 2012) un LVK2018 ir iekļauti pilnīgi visi LVK2013 datu avoti. LVK2018 ir:

- 1) **vispārīgs** – tā avoti netiek ierobežoti atbilstoši konkrētai tēmai, noteiktam valodas funkcionālajam stilam;
- 2) **līdzsvarots** – valodas avoti korpusā ir ievietoti noteiktās proporcijās;
- 3) **mūsdienīgs** – korpusā ir iekļauti tikai mūsdienu rakstu valodas avoti, izņemot Saeimas stenogrammas, kas ir runātas valodas pieraksts (bet ne transkripcija); ar mūsdienu latviešu valodas tekstiem tiek saprasti teksti, kas tapuši pēc 1990. gada;

² „Līdzsvarotais mūsdienu latviešu valodas tekstu korpus” (LVK2018 (bēta)) ir izstrādāts praktiskas ievirzes pētījumu projekta „Daudzslāņu valodas resursu kopa teksta semantiskai analīzei un sintēzei latviešu valodā” (Nr.1.1.1.1/16/A/219) īstenošanas laikā.

³ Pieejams: <http://www.korpus.lv/>.

- 4) **papildināms** – korpusa laika gaitā var tikt papildināts ar jauniem valodas avotiem;
- 5) **reprezentatīvs** – korpusā ir iekļauti gan pilni teksti, gan arī tekstu fragmenti, dažādojot autorus un darbus;
- 6) **anotēts** – korpusa ir automātiski morfoloģiski marķēts (Paikens 2007, 2016; Paikens, Rituma, Pretkalniņa 2013), tam ir pievienoti metadati.

Visi LVK2018 iekļautie datu avoti, kas ir pievienoti šajā korpusa izveides posmā, ir atlasīti automātiski, balstoties uz jaunizstrādāto korpusa datu atlases metodoloģiju. Tāpat korpusa metadati ir pārskatīti un papildināti.

2. LVK2018 uzbūve

Projektējot valodas korpusu, ir jārisina dažādi ar tā uzbūvi saistīti jautājumi – korpusa apjoms, iekļaujamo valodas avotu proporcijas, to tematiskais daudzveidīgums, tekstu atlases kritēriji u.c. –, tāpēc ir ļoti nozīmīgi, lai šāds korpusa tiktu veidots pēc skaidriem uzbūves principiem.

LVK ir veidots, balstoties uz citu valodu vispārīgo korpusu izveides principiem un pieredzi, pielāgojot to latviešu valodas vajadzībām. Tika izmantoti, piem., „Britu Nacionālā korpusa” (*British National Corpus*) (Aston, Burnard 1998; Burnard 2007), „Čehu Nacionālā korpusa” (*Český národní korpus / The Czech National Corpus*) (Čermák 2002; Hnátková, Křen, Procházka, Skoumalová 2014; Křen et al. 2016), „Mūsdienu lietuviešu valodas korpusa” (Kovalevskaitē 2006; Rimkutė et al. 2010) izveides principi.

LVK2018 proporcijas ir noteiktas atbilstoši „Latviešu valodas korpusa koncepcijai” (Levāne-Petrova 2012), laika gaitā tās nedaudz mainot, piem., sadaļas „Populārzinātniski teksti” un „Citi teksti” ir iekļautas „Periodikas” sadaļā, kas ietver šos tekstus. LVK2018 versijā ir piecas sadaļas – daiļliteratūra, periodika, zinātniski teksti, normatīvie akti un Saeimas stenogrammas (sk. 1. attēlu), katru no tām sadalot sīkāk.

1. attēls. LVK2018 proporcijas

Piem., „Periodikas” sadaļā proporcionāli ir iekļauta nacionālo laikrakstu, reģionālo laikrakstu, speciālās periodikas un populārās periodikas daļa (sk. 2. attēlu). LVK2013 versijā populārās periodikas daļa tika dalīta vēl sīkāk, šķirot internetā publicētos izklaides materiālus un drukātos izklaides izdevumus, tomēr šajā korpusa versijas izstrādes posmā datu avoti pēc publicēšanas veida netika dalīti.

2. attēls. LVK2018 periodikas proporcijas

Šobrīd korpusam ir piecas sadaļas, tomēr, papildinot korpusu nākotnē vai veidojot jaunu vispārīgo korpusu, sadaļu skaits varētu būt vēl mazāks, atstājot tikai daiļliterātas, periodikas un t. s. *nedaiļliterātas* (*non-fiction*) sadaļu^{4, 5}. Attiecīgi t. s. *nedaiļliterātas* sadaļā tiktu iekļauti visdažādākie teksti, kas neiekļauj daiļliterātu un periodiku. Korpusa sadaļu skaits ir arī atkarīgs no principiem, pēc kādiem korpusā iekļaujamie datu avoti tiek kategorizēti – publicēšanas veida vai funkcionālā stila. Ja pēc publicēšanas veida, tad, piem., ēdienu receptes, kas ir publicētas dažādos formātos (žurnāls, grāmata, tiešsaistes izdevums), būtu iekļaujamās dažādās korpusa sadaļās.

3. Korpusa tekstu atlases kritēriji

Lai korpus būtu līdzsvarots un tajā pēc iespējas daudzveidīgāk būtu atspoguļota mūsdienu latviešu rakstu valoda, ietverot gan visus valodas funkcionālos stilus, gan arī visplašāko tematiku, ir nepieciešami **korpusa tekstu atlases kritēriji**.

LVK izvēlētos tekstu atlases kritērijus var iedalīt divās grupās:

- 1) tekstu atlases kritēji, kas attiecas uz visu korpusu kopumā;
- 2) kritēriji, kas attiecas uz konkrētu korpusa sadaļu.

⁴ Pieejams: <http://tekstynas.vdu.lt/tekstynas/menu?page=statistics>.

⁵ Pieejams: <http://wiki.korpus.cz/doku.php/en:cnk:syn2000>.

Piem., tekstu atlases kritēriji „Laiks” un „Oriģinālteksti” attiecas uz visu korpusu kopumā, jo korpusā tiek iekļauti tikai tie teksti, kas publicēti kopš 1991. gada un ir oriģinālteksti, nevis tulkojumi. Korpusa veidotāji apzinās, ka periodikas datu avotos varētu būt tulkojumi no citām valodām, piem., ārvalstu ziņas, tomēr šādi teksti netiek īpaši pārskatīti, nosakot to atbilstību iekļaušanai korpusā. Korpusā netiek iekļaut tādi avoti, kas ir neapšaubāmi tulkojumi, piem., daiļdarbu tulkojumi no svešvalodām. Papildinot korpusu, svarīgi arī atlasīt tekstus no kāda konkrēta laikposma, piem., „Periodikas” sadaļa tika papildināta ar 2014.–2016. gadā izdotajiem periodiskajiem izdevumiem. Tādējādi nākamajos korpusa izveides posmos korpusā iekļautie avoti nepārkāpsies, jo tiks ņemti no cita laikposma.

Tā kā korpus nav paredzēts tekstu lasīšanai un lietotājam ir redzama tikai viena konkordances rindīņa, svarīgs ir „Formālas atbilstības kritērijs” – tiek atlasīti teksti, kuros nav, piem., tabulu, matemātisko simbolu u. tml. Ja tomēr attiecīgais tekstu avots ir nozīmīgs, tie fragmenti, kas nav teksts, tiek dzēsti. Visi korpusā iekļautie teksti ir „Autentiski”. Tajos, lai varētu atspoguļot valodu tādu, kāda tā ir, netiek veikti nekādi labojumi. Ja nepieciešamais tekstu avots ir digitalizēts un drukātajā formātā ir bijušas kādas kļūdas, tās tiek atstātas.

Nozīmīgs tekstu atlases kritērijs ir „Pieejamība” – lai korpusu varētu uzkrāt ātrāk un teksti nebūtu jādigitalizē, kas ir laikietilpīgs un dārgs process, tiek atlasīti tie teksti, kas atrodami elektroniskā formātā. Pieejamības kritērijs ir nozīmīgs arī korpusa projektēšanas procesā, jo ne tikai tekstu atlases kritēriji, bet arī pašu valodas avotu pieejamība nosaka to, kādi valodas avoti tiks iekļauti korpusā. Piem., daiļliteratūras sadaļā korpusā tiek iekļauts tas autora daiļdarbs, kas korpusa veidotājiem ir pieejams elektroniskā formātā.

Kritērijs „Tematiskais daudzveidīgums” galvenokārt ir attiecināms uz periodiku un zinātniskajiem tekstiem, tomēr ir pietiekami nozīmīgs arī normatīvo aktu un Saeimas stenogrammu sadaļai. Šis kritērijs nozīmē, ka korpusā jāiekļauj teksti par visdažādākajām tēmām, piem., „Periodikas” sadaļā jāiekļauj raksti par sportu, politiku, ekonomiku u. tml. Tāpat arī jāizvēlas teksti no visdažādākajām zinātņu nozarēm, piem., bioloģijas, valodniecības utt. Saeimas stenogrammas un normatīvie akti ir jāizvēlas pēc iespējas daudzveidīgāki, pārklājot gan dažādas tēmas, gan runātājus. Savukārt tekstu atlases kritērijs „Ģeogrāfija” ir svarīgs reģionālajiem laikrakstiem, kas tiek atlasīti no visiem Latvijas novadiem.

Korpusa izveides iepriekšējos posmos periodikas sadaļā bija svarīgi neiekļaut ziņu par vienu un to pašu notikumu no atšķirīgiem avotiem, piem., dažādos izdevumos dažādu žurnālistu raksti par jaunievēlēto Latvijas prezidentu. Arī šajā korpusa izveides posmā tas ir nozīmīgi, tomēr vēl svarīgāk bija neiekļaut nevienu ziņu vai ziņas daļu, kas varētu dublēties, jo vairums portālu pārpublicē vienas un tās pašas ziņas, turklāt arī dažādos laikposmos. Lai izvairītos no valodas avotu pārklāšanās, visi dokumenti savā starpā tika salīdzināti, rēķinot Breja–Kērtisa (*John Roger Bray, John Thomas Curtis* 1957) nevienlīdzības koeficientu starp divu dokumentu vektoru reprezentācijām (*bag-of-words*). Dažādām korpusa sadaļām tika piemērots atšķirīgs koeficients (sk. tālāk rakstā).

Viens no kopīgajiem kritērijiem, kas izvirzīts, lai nodrošinātu tekstu daudzveidīgumu, – viens teksta vienums nedrīkst pārsniegt 5% no korpusa sadaļas

apjoma. Tādējādi romāni vai promocijas darbi, kas pārsniedz šo apjomu, korpusā pilnā apmērā iekļauti netika, bet tika iekļauti tikai to fragmenti. Minimālais valodas avota apjoms „Periodikas” un „Saeimas stenogrammu” sadaļā ir 30 vārdlietojumi. Piem., BNC (Burnard 2007) ierobežojums viena autora tekstu avotiem ir 120 000 vārdlietojumu, tomēr jāņem vērā, ka BNC apjoms ir 100 miljoni vārdlietojumu, kas ievērojami pārsniedz „Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa” apjomu.

Arī tekstu atlasē kritēriji laika gaitā var tikt pārskatīti un mainīti, piem., savulaik korpusā tika iekļauti tikai drukāti avoti, toties tagad, kad vairums ziņu u. c. valodas avotu ir tikai elektroniskā formātā, šis kritērijs vairs nav svarīgs.

Tālākajās apakšnodaļās ir aplūkoti sīkāki principi, pēc kādiem katra konkrētā korpusa sadaļa tika aizpildīta ar tekstiem.

3.1. Periodika

Sadaļa „Periodika” tiek iedalīta vairākās daļās: nacionālā, reģionālā, populārā un speciālā periodika (sk. 2. attēlu). Iepriekšējos korpusa izveides posmos bija svarīgi šķirt nacionālos laikrakstus un interneta ziņas (ziņas drukātā un elektroniskā formātā, attiecīgi pēc publicēšanas veida), bet šajā korpusa izveides posmā šīs apakšsadaļas tika apvienotas, jo gan drukātie plašsaziņas līdzekļi, gan arī daudzi nacionālie elektroniskie plašsaziņas līdzekļi parasti nacionālās ziņas pārpublicē no nacionālās informācijas aģentūras LETA, reizēm tikai nedaudz pamainot saturu. Tādējādi, lai ziņas nepārklātos, teksti ņemti tikai no aģentūras LETA arhīva.

Arī reģionālā, populārā un speciālā periodika tiek atlasīta no aģentūras LETA monitoringa resursiem, kas ietilpst konkrētajās kategorijās. Reģionālās periodikas apakšsadaļā ir iekļauti raksti no *rekurzeme.lv*, *ezerzeme.lv*, *edruva.lv* u. c. reģionālajiem ziņu izdevumiem, populārās periodikas apakšsadaļā ir iekļauti raksti no *cosmo.lv*, *slavenibas.lv* u. tml., savukārt speciālās periodikas apakšsadaļā ir iekļauti dažādi raksti no *kroders.lv*, *cetrassezonas.lv*, *garaza.lv* u. c.

Lai nodrošinātu atlasīto datu tematisko daudzveidību, piem., par politiku, sportu, izklaidi, tika atlasītas visas ziņas no viena avota, kas ir publicētas vienā datumā, tādējādi nodrošinot, ka datus tiek saglabāta atbilstīgā proporcija. Tāpat tematiskās daudzveidības nodrošināšanai dati tiek atlasīti vienmērīgi no visa pieejamā laikposma, piem., no 2014. gada līdz 2016. gadam. Lai izvairītos gan no nejaušas tematu pārklāšanās, gan arī no teksta pārklāšanās kopumā, visi dokumenti savā starpā tiek salīdzināti, rēķinot Breja–Kērtisa nevienlīdzības koeficientu starp divu dokumentu vektoru reprezentācijām. Ja nevienlīdzības koeficients ir mazāks par 0,6, tad no šiem dokumentiem tiek paturēts tikai viens. Analizējot atlasītos datus, secināts, ka vairākos dokumentos ir sastopamas dažādas kvalitātes problēmas, piem., pazudušas atstarpes, pieturzīmes un izkropļoti simboli. Lai atlasītu tikai kvalitatīvus tekstus, resp., korpusam un pētniecībai atbilstošus tekstus, no periodikas datiem tika izgūts saraksts ar tādiem vārdiem, kas nav iekļauti latviešu valodas vārdnīcās, nav vispārzināmi un lietojami vārdi, piem., vārdi svešvalodās u. tml. Ja dokumentā bija vārdi, kas ir sarakstā, tad dokuments netika iekļauts korpusā. Tāpat kvalitātes nodrošināšanai tika noteikta augstākā un zemākā procentuālā robeža vārdu skaitam pret visiem vienumiem dokumentā, par vienumiem uzskatot vārdus, skaitļus, pieturzīmes un citus simbolus. Procentuālās

robežas tika noteiktas, balstoties uz iepriekšējā korpusa LVK2013 versijas manuāli pārbaudītajiem periodikas datiem, proti, no 64% līdz 88%. Augšējā robeža tika noteikta, lai izvairītos no dokumentiem, kuros kaut kādu iemeslu dēļ bija pārāk maz interpunkcijas zīmju vai to nebija vispār. Dokumenti tika atlasīti, ievērojot arī vārdu skaita limitu. Periodikas sadaļā dokumentā ir jābūt vismaz 30 vārdiem, bet ne vairāk kā 2500 vārdiem.

3.2. Daiļliteratūra

LU MII datubāzē ir pieejami dažādi daiļliteratūras avoti, kas atbilstošās proporcijās ir iekļauti LVK2018. Ja daiļdarbs pārsniedz noteikto datu avota apjomu („Daiļliteratūras” sadaļā 55 000 vārdlietojumu), tad, lai konkrētais autora darbs nedominētu korpusā, tika iekļauts tikai tā fragments. Veidojot daiļliteratūras sadaļu, nozīmīgs ir pieejamības kritērijs – resp., tiek iekļauti tie daiļliteratūras avoti noteiktās proporcijās, kas ir pieejami LU MII datubāzē elektroniskā formātā.

3.3. Zinātne

Zinātniskie teksti tika atlasīti no internetā pieejamajiem promocijas darbiem vai to kopsavilkumiem. Lai nodrošinātu teksta daudzveidību, promocijas darbi vai to kopsavilkumi atlasīti no dažādām zinātņu nozarēm, piem., fizikas, valodniecības, ģeogrāfijas utt. (sk. 3. attēlu).

3. attēls. LVK2018 zinātnisko tekstu proporcijas

Lai arī visas pārējās korpusa sadaļas tika aizpildītas ar automātiski atlasītiem datu avotiem, zinātnisko tekstu sadaļā datu atlasē process notika pusautomātiski. Sākotnēji manuāli tika atlasīti korpusam formāli atbilstoši promocijas darbi (formālās atbilstības kritērijs) no konkrētām zinātņu nozarēm, tad tika veikta to pusautomātiska apstrāde. Tā kā dokumenti ir pieejami PDF formātā un no tiem ir nepieciešams izgūt tekstu, tie vispirms tika pārveidoti HTML formātā, kurā tika veikta tālāka apstrāde. HTML dokuments tika automātiski attīrīts no korpusā neiekļaujamās informācijas – tabulām, attēliem, lapaspušu galvenēm un kājenēm.

3.4. Normatīvie akti

LVK2018 „Normatīvo aktu” sadaļā tika iekļauti Latvijas Republikas Saeimas pieņemtie likumi, kas automātiski atlasīti no vietnes *likumi.lv*.

Lai nodrošinātu iespēju korpusu papildināt arī turpmāk, izvairoties no datu pārklāšanās, tika atlasīti tikai tie likumi, kas ir pieņemti laika posmā no 01.01.2014. līdz 31.12.2016. Tāpat, lai izvairītos no datu pārklāšanās, likumu grozījumi korpusā netika iekļauti. Korpusam netika izvēlēti arī protokoli un citi normatīvie akti, kuru tīmekļa lapās ir paralēli atrodami to tulkojumi svešvalodās, jo, pirmkārt, korpusā teksti svešvalodās nav iekļaujami, bet, otrkārt, šādi katrs automātiski atlasītais tekstu fails būtu manuāli jāapstrādā – jāizņem teksts svešvalodā.

Katrs likums satur ļoti līdzīgu galveni un kājeni, un visi likumi sākas ar tekstu „Saeima ir pieņēmusi un Valsts prezidents izsludina šādu likumu:”, bet beidzas ar datumu, kad tas ir pieņemts, kad stājas spēkā, atbildīgā parakstu un datumu. Lai saglabātu korpusa tekstu daudzveidīgumu, šī informācija tika atņemta un paturēts tikai likuma pamatteksts. Papildus tiek apkopoti metadati – likuma pieņemšanas datums, spēkā stāšanās datums un oficiālās publikācijas numurs.

Visi likumi no konkrētā laikposma tiek automātiski saglabāti datubāzē, no kuras savukārt korpusam tiek atlasīti pilni dokumenti, kas nepārsniedz noteikto vārdlietojumu ierobežojumu (22 000 vārdlietojumu „Normatīvo aktu” sadaļai). Tomēr jāņem vērā, ka korpusā ir tikai 10 miljonu vārdlietojumu liels un labāk ir iekļaut daudzus dokumentus ar mazāku vārdlietojumu skaitu, nevis vienu lielu dokumentu, tādējādi nodrošinot leksikas daudzveidību. Lai nodrošinātu sadaļas tematisko daudzveidīgumu, dokumenti tika atlasīti vienmērīgi gan no visa savāktā perioda, gan atbilstoši tematikai. Papildus tematiskā daudzveidīguma nodrošināšanai, līdzīgi kā pie periodikas atlasē, korpusā tika iekļauts viens no diviem dokumentiem, ja starp tiem Breja–Kērtisa nevienlīdzības koeficients bija mazāks par 0,65. Normatīvajos dokumentos bieži ir sastopamas standartfrāzes. Lai tās nedominētu biežumu sarakstā, korpusā tika iekļauts tikai viens no diviem dokumentiem, ja šajos dokumentos bija kopīga frāze vismaz 12 vārdu garumā.

3.5. Saeimas stenogrammas

Saeimas stenogrammas tika atlasītas no Saeimas stenogrammu korpusa (<http://saeima.korpus.lv>). Saeimas stenogrammu korpusā par vienu dokumentu tiek uzskatīta viena runātāja viens nepārtraukts uzrunas fragments no tribīnes. Tāpat viens Saeimas stenogrammu dokuments tika definēts arī LVK2018.

Saeimas stenogrammu korpusā runātāji ir iedalīti sešās kategorijās – deputāti, Latvijas institūciju pārstāvji, ministriju amatpersonas, Saeimas amatpersonas,

valdība un ārvalstu viesi. No šīm kategorijām korpusam tika pievienoti dokumenti tikai no deputātu uzrunām, jo citās kategorijās var ietilpt tulkota runa, piem., ārvalstu viesi, turklāt citās kategorijās mēdz būt pārāk vienvēidīgi teksti, piem., Saeimas amatpersonu teiktajā.

No Saeimas stenogrammu korpusā marķētajiem sēžu tipiem – kārtējā, noslēguma, svinīgā, ārkārtas vai ārkārtas sesijas – tika atlasītas uzrunas ar tipu „kārtējā”, jo citos sēžu tipos mēdz parādīties ļoti specifiski teksti, kas var negatīvi ietekmēt LVK2018 kopējo statistiku, piem., deputātu svinīgi zvēresti, uzsākot darbu Saeimā.

Lai atvieglotu korpusa tālāku papildināšanu nākotnē, dokumenti tika vienmērīgi atlasīti no 01.01.2014. līdz 31.12.2016. No katra datuma tika iekļauts vienāds skaits īsāko uzrunu. Arī šajā korpusa sadaļā, tāpat kā „Periodikas” sadaļā, dokumentā ir jābūt vismaz 30 vārdiem.

4. Metadati

LVK2018 ir anotēts – tam ir izstrādāta metadatu kopa. Metadati jeb „dati par datiem” ir informācija, kas apraksta kādu citu datu kopu. Līdzīgi kā ar korpusa proporcijām un tekstu atlasē kritērijiem, arī metadatu kopa jaunajai korpusa versijai LVK2018 tika pārskatīta un atjaunināta. Daļa no metadatu kopas vērtībām lietotājiem bija redzamas, piem., žanrs, avota autors u.c., savukārt citas vērtības bija paredzētas tikai iekšējai lietošanai, piem., vārdlietojumu skaits, atslēgvārdi u.c. Jaunajā metadatu kopā paturētas tikai tās vērtības, kas ir nosakāmas un lietotājiem nepieciešamas, piem., avota autora dzimums vairs netiks iekļauts, jo vairumā gadījumu tas nav zināms.

LVK2018 ir publiski redzami trīs metadatu lauki – **unikālais identifikators** (id), **sadaļa** un **atsauce**. Atsauces formāts ir atkarīgs no dokumenta sadaļas:

- 1) normatīvajiem aktiem – *{nosaukums}, pieņemts {pieņemšanas datums}, publicēts {publicēšanas datums}*;
- 2) Saeimas stenogrammām:
 - a) LVK2018 versijas datu avotiem: *No Saeimas stenogrammām {datums}, {runātājs} ({politiskā frakcija})*;
 - b) LVK2013 versijas datu avotiem: *No Saeimas stenogrammām {datums}, temats: “{temats}”, {avota URL}*;
- 3) zinātnei – *{autors}, {nosaukums} ({zinātnes nozare}), {gads}*;
- 4) daiļliteratūrai – *{autors}, {darba nosaukums} ({nodaļas nosaukums}), {izdošanas vieta}, {izdevniecība}, {izdošanas gads}*;
- 5) periodikai – *{nosaukums}, {avots}, {publicēts}, {apakšsadaļa}*.

Parasti pētījumos korpuss tiek izmantots viss kopumā, tāpēc arī publiski redzamie metadati iekļauj tikai iepriekš redzamos laukus. Piem., korpusa lietotājs varēs uzzināt, ka konkrētais avots ir no „Periodikas” sadaļas, raksta nosaukumu, avotu, izdošanas datumu un apakšsadaļu, kurai konkrētais datu avots pieder, tāpat ir redzams unikālais ID, kam redzamais burts p arī norāda uz piederību konkrētajai sadaļai – periodikai (sk. 4. attēlu).

ID p5092

Atsauce Stikla pulēšana, Latvijas Avīze, 2011, nacionālā periodika.

Sadaļa Periodika

4. attēls. Periodikas metadati

5. Morfoloģiskais marķējums

LVK2018 ir automātiski morfoloģiski marķēts, izmantojot LUMII Mākslīgā intelekta laboratorijā izstrādāto morfoloģisko marķētāju (Paikens 2007; Paikens 2016; Paikens, Rituma, Pretkalniņa 2013). Morfoloģiski marķētā korpusā katram vārdlietojumam ir automātiski pievienotas tam atbilstošas morfoloģiskās pazīmes⁶ un pamatforma (lemma). Kā jau tika minēts raksta ievadā, korpusa morfoloģiskais marķējums nedaudz atšķiras no tradicionālajiem gramatikas aprakstiem (Ahero et al. 1959; Nītiņa, Grigorjevs 2013), piem., korpusa morfoloģiskais marķējums neiekļauj noliegtos vietniekvārdus kā atsevišķu vietniekvārdu grupu, jo noliegumam ir paredzēta īpaša morfoloģiskā pazīme, tāpat neiekļauj instrumentāli kā atsevišķu locījumu, jo tam nav īpašu formālu pazīmju un tā statuss ir diskutējams (Smiltnece 2013, 352–353), turklāt neviena korpusa vienība nevar būt bez marķējuma (sk. 5. attēlā redzamos nenoteiktos vietniekvārdus, par kuru noliegumu liecina 7. pazīme – y).

First | Previous Page 3 of 3,757 Go Next | Last

1 Saeimas st...	dažādi novadu apvienošanas modeļi . Bet tam noteikti nebija	nekāda /piOfsny	plašāka ekonomiskā pamatojuma , un diez vai to var raksturot	
2 Saeimas st...	pašvaldību sistēmas jeb novadu robežas . Otrkārt , noteikti ir	daudzli /pi0mpn_	citi steidzamāki un svarīgāki uzdevumi , kas veicami politiķiem	
3 Saeimas st...	pašvaldību sistēmas jeb novadu robežas . Otrkārt , noteikti ir daudzi	citi /pi0mpnn	steidzamāki un svarīgāki uzdevumi , kas veicami politiķiem	
4 Saeimas st...	kas veicami politiķiem un valsts pārvaldei . Ir krietni daudz	ko /pi000an	darīt uzņēmējdarbības vides uzlabošanā , darba vietu radīšanā	
5 Saeimas st...	atbalstā un tā tālāk . Es noteikti varētu nosaukt vismaz simts	citus /pi0mpan	šobrīd valstiski svarīgākus un steidzamākus darbus , kas veicami	
6 Saeimas st...	Zemnieku savienība to ļoti labi atceras un piemin to ... ka ļoti	daudzus /pi0mpa_	likumprojekts , kurus , piemēram , mēs iesniedzam un kurus	
7 Saeimas st...	parlamentā . Un , kolēģi , tādi likumprojekti mums parlamentā ir	vairāki /pi0mpn_	. Nu , piemēram , mums ir kopīgi iesniegts projekts attiecībā	
8 Saeimas st...	komisijā viņi paskata to , taču jautājums netiek virzīts tālāk un	nekāds /pi0msny	lēmums pieņemts netiek . Pa to laiku īstenībā banku turpina	
9 Saeimas st...	komisijas vadītājam , vai koalīcijai , bet gan prioritārā secībā ,	kādā /pi0fsln	tie nāk , piemēram , vai nu no Ministru kabineta , vai ar Saeimas	
10 Saeimas st...	militāro transportu ar Krievijas armijas numurziņiem . Tikpat kā	nevienam /pi0msdy	militārajam ekspertam vaļs nav šaubu , ka iebrukumu Krimā	
11 Saeimas st...	vajāšanu , un līdz ar to arī šis Krievijas arguments neiztur	nekādu /pi0fsay	kritiku . Arvien vairāk šķiet , ka Krievija ir pārrēķinājusi	
12 Saeimas st...	pretējo . Arguments par Krimas pašsārdzības vienībām neiztur	jebkādu /pi0fsan	kritiku un kļūst par absurdu , kad ikviens medijos var aplūkot	
13 Saeimas st...	Krieviju un Ukrainas valdību par mērīgu krīzes noregulējumu , pie	kā /pi000gn	patlaban strādā Rietumvalstu un Krievijas amatpersonas un diplomāti	
14 Saeimas st...	NATO gaisa patrulēs Baltijas valstu gaisa telpā . Tas ir fons ,	kādā /pi0fsln	arī Eiropas lietu komisija ir gatavojusi savu kārtējo paziņojumu	
15 Saeimas st...	situācijas attīstībai Ukrainā . Un atļaujiet atgādināt tikai	dažus /pi0mpan	iepriekšējos lēmumus . Piemēram , vēl pirms 20. februāra traģiskajiem	
16 Saeimas st...	armijas vienību izvešanu uz to bāzēm saskan ar to pozīciju ,	kādu /pi0msan	Latvijas ārlietu ministrs puda iepriekš minētajā Eiropas Savienības	
17 Saeimas st...	padomes ārkārtas sanāksmē , kā arī ar Latvijas oficiālo nostāju ,	kādu /pi0msan	Latvijas premjerministre Laimdota Straujuma šodien paudis Briselē	
18 Saeimas st...	... šie tirgotāji ir nolīguši advokātus , kuriem var zvanīt	jebkurā /pi0fsln	diennakts laikā un saņemt norādījumus . Pašreizējā situācija	

5. attēls. LVK korpusa saskarne, meklējot nenoteiktos vietniekvārdus

Morfoloģiskais marķējums iekļauj 14 marķēšanas kategorijas, marķējot gan tradicionālās vārdšķiras, piem., lietvārdu, vietniekvārdu, skaitļa vārdu utt., gan arī īpaši definētas papildgrupas, piem., pieturzīmes, saīsinājumus u.c. Katrai marķēšanas kategorijai atbilst īpašas marķējuma pazīmes, kas redzamas tālākajos apzīmējumos. Piem., vārdformai *pavasara* ir pievienots šāds morfoloģiskais marķējums [ncmsg2], kas ir atšifrējams kā n – lietvārds, c – sugasvārds, m – vīriešu dzimte, s – vienskaitlis, g – ģenitīvs, 2 – 2. deklinācija.

⁶ LVK2018 morfoloģisko pazīmju kopa ir atrodamā latviešu valodas korpusu mājaslapā: www.korpus.lv.

Šāds marķējums ļauj ievērojami ātrāk un vieglāk tekstu korpusā atrast nepieciešamos datus pēc konkrētām pazīmēm, piem., visus 6. deklinācijas lietvārdus, izmantojot īpašu vaicājumu [tag = „n...6”] (sk. 6. attēlu).

The screenshot shows the NoSketch Engine search interface. The search bar contains 'LVK2018 (beta)'. The results list shows 25 items, all starting with 'Saeimas st...'. The results are displayed in a table-like format with columns for the item number, the text snippet, and a small icon. The text snippets are related to the Latvian parliament (Saeima) and its activities.

6. attēls. LVK2018 korpusa saskarne, meklējot 6. deklinācijas lietvārdus

Jāņem vērā, ka latviešu valoda ir morfoloģiski daudznozīmīga un automātiskās marķēšanas procesā reizēm nav iespējams atšķirt konkrētas vārdformas vārdšķiru, tāpēc dažkārt novērojamas kļūdas morfoloģiskajā marķējumā, jo īpaši formu homonīnijas gadījumos. Piem., 6. attēlā ir redzams, ka vīriešu dzimtes lietvārda *grants*⁷ daudzskaitļa nominatīvs *granti* automātiski ir marķēts kā sieviešu dzimtes vienskaitļa akuzatīvs un tiek piedāvāts kā 6. deklinācijas lietvārds.

6. Korpusa pieejamība un meklēšanas iespējas

Korpuss ir pieejams, izmantojot korpusa platformu *NoSketch Engine* (Rychlý 2007) vietnē: <http://nosketch.korpuss.lv/> (sk. 5., 6. un 7. attēlu ar korpusa saskarni). Lai varētu veiksmīgi strādāt ar korpusu un tajā atrast sev nepieciešamos datus, korpusa platformā ir jābūt dažādām datu meklēšanas, skatīšanas un kārtīšanas iespējām.⁸

Korpusa pamatā ir vienkāršā meklēšana. Tas nozīmē, ka korpusā var meklēt konkrētu vārdformu (tā jāievada laukā „Vienkāršs vaicājums” vai laukā „Vārdforma”), vārdu pamatformā un visas tā formas (pamatforma jāievada

⁷ 1. Finansējums, ko piešķir kāda darba veikšanai vai speciālam mērķim un kas nav jāatdod (www.tezaurs.lv).

⁸ Detalizēta korpusa lietošanas instrukcija „LVK2018: meklēšana korpusā” ir atrodama latviešu valodas korpusu mājaslapā: www.korpuss.lv, kā arī vietnē: http://www.clarin.lv/attachments/LVK2018_meklesana.pdf.

laukā „Vienkāršs vaicājums” vai „Pamatforma”). Piem., ievadot vaicājumu *iet*, tiek atrastas visas verba *iet* vārdformas *ēju, iesim, gāju, gāja* utt. Var meklēt arī vārdu savienojumus. Ja laukā „Frāze” tiks ierakstīts konkrēts vārdu savienojums, piem., *lasīt grāmatu, iet uz skolu*, tiks atrasts vārdu savienojums tieši šajā formā, savukārt, ja laukā „Vienkāršs vaicājums” ieraksta vārdu savienojuma komponentus pamatformā (*lasīt grāmatu*), tie tiks atrasti dažādās formās: *lasot grāmatu, lasīju grāmatu, lasīdama grāmatu, lasīsim grāmatas* utt. Korpusā var arī meklēt, izmantojot regulārās izteiksmes⁹ – pēc simboliem, kas definē meklējamo izteiksmi. Piem., ja laukā „Pamatforma” vai „Vārdforma” meklē *vasar**, tiks atrasti visas vārdformas, kas sākas ar *vasar* (*vasara, vasaras, vasarnīcas* utt.).

Ja ir nepieciešama specifiskāka informācija, korpusā iespējams arī meklēt ar sarežģītākiem vaicājumiem. Korpusā var meklēt pēc morfoloģiskajām pazīmēm, piem., var atrast visus ciešamās kārtas tagadnes divdabjus, izmantojot īpašu vaicājumu [tag=„vmnpd...ppnp.”]. Tāpat korpusā var meklēt, izmantojot pamatformu un/vai morfoloģiskās pazīmes. Piem., ja laukā „CQL” ievada vaicājumu [lemma=„tikt”][tag=„vmnpd...psnp.”] (palīgdarbības vārda *tikt* pamatforma + ciešamās kārtas pagātnes divdabja morfoloģiskās pazīmes), var atrast ciešamās kārtas vienkāršo laiku formas (sk. 7. attēlu). Izmantojot visdažādākās meklēšanas vaicājumu kombinācijas, korpusā ir iespējams atrast nepieciešamās konstrukcijas, nemaz nenorādot konkrētu(-s) vārdu(-s), piem., darbības vārds vajadzības izteiksmē + lietvārds, palīgdarbības vārds + ciešamās kārtas tagadnes divdabjs utt.

kas ietver likumu ... likuma normas . Tātad visa procedūra , kā	tiek /vtnipi130an	pieņemti /vmnpdmpnpsnpn	lēmumi Latvijā , lēmumu sadalīšana , kompete
ia , ka nevar būt pārāk mazi granti , tāpēc granti ir jāapvieno .	tika /vtnsi130an	apvienoti /vmnpdmpnpsnpn	kopā divi granti - no viena nāca kosmoss , no o
viens no profesoriem , kuri ... varbūt viens no pirmajiem , kuri	tika /vtnsi130an	uzaicināti /vmnpdmpnpsnpn	uz ārzemēm - uz Rietumvāciju . Dzīvoju Kaizer
vai tas ir tikai deklaratīvs uzstādījums vai kā absolūts princips	tiek /vtnipi130an	realizēts /vmnpdmsnpsnpn	arī katrā veiktajā darbā ? Ziņojumā apliecināts
pēc nepilna gada tuvdzīves dēļ neskaitāmu miljonu apmērā	tiek /vtnipi130an	apzagta /vmnpdmsnpsnpn	valsts un tās iedzīvotāji ? Taču tas ir tikai vien-
na veidošanai , secinot , ka Kremļa politikai labvēlīga situācija	tiek /vtnipi130an	panākta /vmnpdmsnpsnpn	, cieši sadarbojoties prokrieviskajām politiskaj
iljoniem eiro katru dienu . Arī termiņuzturēšanās atļaujas , kas	tiek /vtnipi130an	nekontrolēti /vmnpdmpnpsnpn	piešķirtas pret ieguldījumiem nekustamajā īpa
eguvumi , kuri saistīti pirmām kārtām ar to , ka parastās tiesas	tiek /vtnipi130an	atslogotas /vmnpdmpnpsnpn	un tajās netiek izskatītas lietas , par kurām at
tvijā ir tikai divas šķīrējtiesas , kuras atbilst kritērijiem , kuras	tiek /vtnipi130an	izceltas /vmnpdmpnpsnpn	saskaņā ar piedāvātā likumprojekta redakciju
ta jautājums . Bet par to , ka konkrēti šīs šķīrējtiesas intereses	tiek /vtnipi130an	lobētas /vmnpdmpnpsnpn	, skatoties caur konkrētā likumprojekta prizmi
ru kabinetā , lai tur vēlreiz notiktu diskusija pēc būtības un lai	tiktu /vtnc0i100an	iestrādāti /vmnpdmpnpsnpn	tādi kritēriji , atbilstoši kuriem šķīrējtiesas nā
idēļ šāda komisija savulaik netika izveidota . Uz komisijas sēdi	tika /vtnsi130an	uzaicināti /vmnpdmpnpsnpn	eksperti , kā arī Vides aizsardzības un reģionāl
skatīties , ka patiešām šīs lietas , runājot par valsts attīstību ,	tiek /vtnipi130an	risinātas /vmnpdmpnpsnpn	kompleksi . Un līdz ar to komisijas vārdā es aic
jaunas tehnoloģijas : panāk to , ka tā gāze , kas atrodas iezos ,	tiek /vtnipi130an	sadrupināta /vmnpdmpnpsnpn	ar dažādām tehnoloģijām , metodēm . Nāk ārē
nportēšanas sistēmas uz eksportēšanas sistēmu . Bet tas lēnām	tiek /vtnipi130an	darīts /vmnpdmsnpsnpn	. Bet Amerika nesteidzas . Kāpēc ? Tāpēc , ka .
viņi , deviņi plus viens - un atkal no jauna šīs skaitļu virtenes ir	tikušas /vtnpdmpnpsnpn	apspriestas /vmnpdmpnpsnpn	un skandētas kopš tālā 1993. gada , kad tika p
s ir tikušas apspriestas un skandētas kopš tālā 1993. gada , kad	tika /vtnsi130an	pieņemts /vmnpdmsnpsnpn	pirmais lēmums par nepieciešamību īstenot ad
sma risinājums . Tagad būs skaidrs , ka tieši plānošanas reģioni	tiek /vtnipi130an	atzīti /vmnpdmpnpsnpn	par optimālu pārvaldes formu reģionālā līmenī
rid novājinātājam sabiedriskajām organizācijām jeb biedrībām	tika /vtnsi130an	izveidoti /vmnpdmpnpsnpn	plānošanas reģioni , vienojoties par to padome
ecājos , ka beidzot neskaidrība par plānošanas reģionu nākotni	tiks /vtnifi130an	izbeigta /vmnpdmsnpsnpn	ar šo likumu . Plānošanas reģioni ir palīdzējuši

7. attēls. LVK2018 sarkarne ar meklēšanas vaicājumu [lemma=„tikt”][tag=„vmnpd...psnp.”]

Korpusa platformā ir ne tikai dažādas meklēšanas iespējas, bet arī dažādas datu skatīšanas un kārtotošanas iespējas. Lai iespējami ērtāk aplūkotu atrastos

⁹ Detalizēta instrukcija par regulārajām izteiksmēm un to lietošanu korpusā ir atrodamā latviešu valodas korpusu mājaslapā: www.korpuss.lv, kā arī vietnē: http://www.clarin.lv/attachments/27042018reg_izt.pdf.

korpusa datus, var pielāgot nepieciešamā konteksta garumu, lapā redzamo piemēru skaitu, labās un kreisās puses apkaimes apjomu. Savukārt datus var kārtot pēc kreisās/labās puses konteksta, pēc dokumenta nosaukuma u. c. iespējām. Tāpat arī var veidot atrasto datu biežuma sarakstus, kā arī visbeidzot saglabāt atrastos datus, lai ar tiem varētu ērti strādāt sev ērtā laikā. Lai arī korpusa lietotājs var izvēlēties konteksta garumu u. c. korpusa parametrus, korpusā nekad nevarēs izlasīt visu iekļauto avotu, piem., daļdarbu, bet tikai īsu tā fragmentu.

7. Kā atsaukties uz LVK?

Laika gaitā latviešu valodai ir izstrādāti dažādi korpusi (vispārīgais korpus ar tā versijām, runas korpusi, specializētie korpusi, paralēlie korpusi u. c.), tāpēc ir jo īpaši svarīgi precīzi atsaukties uz izmantoto korpusu, minot tā nosaukumu, akronīmu, ja tāds ir, un citu pieejamo informāciju par korpusu. Uz LVK materiāliem ir jāatsaucas:

„Līdzsvarotais mūsdienu latviešu valodas tekstu korpus” – LVK2018 (bēta).
LU Matemātikas un informātikas institūts, 2018. Pieejams: www.korpuss.lv.

8. Secinājumi un nākotnes ieceres

LVK jau vairāk nekā desmit gadus ieņem stabilu vietu latviešu valodas izpētē, tāpēc jauna, paplašināta LVK versija ir ļoti gaidīta un nepieciešama. Lai arī mūsdienās vispārīgais valodas korpus 10 miljonu vārdlietojumu apmērā nav liels, tomēr tas ir gan vispārīgu gramatikas pētījumu, gan arī noteiktu semantikas pētījumu pamats, piem., verbu valenču pētījumi (Nešpore, Saulīte, Grūzītis, Garkāje 2012; Kuplā 2013; Ciematniece 2015). Tāpat korpus tiek izmantots dažādos lingvistiskajos pētījumos un ikdienas mācību procesā. LVK daļa tiek arī sintaktiski un semantiski marķēta un būs pamats jauniem latviešu valodas gramatikas pētījumiem.

LVK jaunā versija ir ne tikai lielāks vispārīgais latviešu valodas korpus, bet arī jauns solis vispārīga korpusa izstrādē latviešu valodai, jo šajā korpusa izstrādes posmā LVK tika pārskatīti un standartizēti daudzi korpusa izstrādes principi. Piem., tika standartizēti metadatu lauki un vērtības, kā arī izstrādāta korpusa datu atlases metodoloģija, bet pats korpus ieguva jaunu tīmekļa platformu. Ar jaunizstrādāto datu atlases metodoloģiju korpusa avoti tika atlasīti gandrīz pilnīgi automātiski. Tas nākotnē ļautu korpusu papildināt līdz t. s. mūsdienu vispārīgā korpusa standartam – 100 miljoniem vai vairāk vārdlietojumu, kas būtu arī leksikogrāfijas un citu pētījumu pamats.

Saīsinājumi un apzīmējumi

BNC	<i>British National Corpus</i>
LU MII	Latvijas Universitātes Matemātikas un informātikas institūts
LVA	Latviešu valodas aģentūra
LVK	<i>Līdzsvarotais latviešu valodas tekstu korpus</i> . Pieejams: www.korpuss.lv
LVK2013	<i>Līdzsvarotā latviešu valodas tekstu korpusa 2013. gada versija</i> . Pieejams: http://www.korpuss.lv/id/LVK2013

- LVK2018 *Līdzsvarotā latviešu valodas tekstu korpusa* 2018. gada versija. Pieejams: <http://www.korpuss.lv/id/LVK2018>
- VVA Valsts valodas aģentūra

Literatūra

1. Ahero, Antonija et al. 1959. *Mūsdienu latviešu literārās valodas gramatika*. 1. *Fonētika un morfoloģija*. Rīga: LPSR Zinātņu akadēmijas izdevniecība.
2. Andronova, Everita, Andronovs, Aleksejs. 2011. Latviešu valodas korpus un tā izmantošana. *Valodas prakse: vērojumi un ieteikumi*. 6. Rīga: Latviešu valodas aģentūra, 41–57.
3. Aston, Guy, Burnard, Lou. 1997. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press. Pieejams: <http://corpus.leeds.ac.uk/teaching/aston-burnard-bnc.pdf>.
4. Burnard, Lou (ed.). 2007. *Reference Guide for the British National Corpus (XML Edition)*. Published for the British National Corpus Consortium by Oxford University Computing Services. Pieejams: <http://www.natcorp.ox.ac.uk/docs/URG/>.
5. Bussmann, Hadumod. 2006. *Routledge Dictionary of Language and Linguistics*. London: Taylor & Francis. Pieejams: https://www.e-reading.club/bookreader.php/142124/Routledge_Dictionary_of_Language_and_Linguistics.pdf.
6. Bray, J. Roger, Curtis, John Thomas. 1957. An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs*. 27(4), 325–349.
7. Ciematniece, Līga. 2015. *Teikuma semantiskās un sintaktiskās struktūras attieksmes mūsdienu latviešu valodā*. Promocijas darbs filoloģijas doktora grāda iegūšanai valodniecības zinātņu nozares latviešu sinhroniskās valodniecības apakšnozarē. Rīga: Latvijas Universitāte.
8. *Collins English Dictionary*. 13th edition. Glasgow: HarperCollins Publishers Ltd., 2018. Pieejams: <https://www.collinsdictionary.com/dictionary/english/>.
9. Čermák, František. 2002. Today's corpus linguistics. Some open questions. *International Journal of Corpus Linguistics*. 7(2). Amsterdam: John Benjamins, 265–282.
10. Hnátková, Milena, Křen, Michal, Procházka, Pavel, Skoumalová, Hana. 2014. The SYN-series corpora of written Czech. *Proceedings of LREC 2014*. Reykjavík: ELRA, 160–164. Pieejams: http://www.lrecconf.org/proceedings/lrec2014/pdf/294_Paper.pdf.
11. Kovalevskaitė, Jolanta. 2006. Dabartinės lietuvių kalbos tekstynas – 10 metų kaupimo ir naudojimo patirtis. *Prace Baltyszne*. 3. *Język. Literatura. Kultura*. Warszawa: Uniwersytet Warszawski, 231–241.
12. Křen, Michal, Cvrček, Václav, Čapka, Tomáš, Čermáková, Anna, Hnátková, Milena, Chlumská, Lucie, Jelinek, Tomáš, Kováříková, Dominika, Petkevič, Vladimír, Procházka, Pavel, Skoumalová, Hana, Škrabal, Michal, Truneček, Petr, Vondříčka, Pavel, Zasina, Adrian Jan. 2016. SYN2015: Representative corpus of contemporary written Czech. *Proceedings of LREC 2016*. Portorož: ELRA, 2522–2528. Pieejams: http://www.lrec-conf.org/proceedings/lrec2016/pdf/186_Paper.pdf.
13. Kuplā, Ieva. 2013. *Kustības verbu daudznozīmība latviešu valodā*. Promocijas darbs filoloģijas doktora grāda iegūšanai valodniecības zinātņu nozares latviešu sinhroniskās valodniecības apakšnozarē. Rīga: Latvijas Universitāte.

14. Levāne-Petrova, Kristīne. 2011. Morfoloģiski marķēta valodas korpusa izmantošana valodas izpētē. *Vārds un tā pētīšanas aspekti*. 15(1). Liepāja: LiePA, 187–193.
15. Levāne-Petrova, Kristīne. 2012. Līdzsvarots mūsdienu latviešu valodas tekstu korpusss un tā tekstu atlases kritēriji. *Baltistica*. VIII. *Priedas*. Vilnius: Vilniaus Universiteto leidykla, 89–98.
16. Nešpore, Gunta, Saulīte, Baiba, Grūzītis, Normunds, Garkāje, Ginta. 2012. Towards a Latvian Valency Lexicon. *Human Language Technologies – The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012*. 247. *Frontiers in Artificial Intelligence and Applications*. Tavast, Arvi, Muischnek, Kadri, Koit, Mare (eds.). Amsterdam: IOS Press, 156–161.
17. Nītiņa, Daina, Grigorjevs, Juris (red.). 2013. *Latviešu valodas gramatika*. Rīga: LU Latviešu valodas institūts.
18. Paikens, Pēteris. 2008. Lexicon-based morphological analysis of Latvian language. *Proceedings of the 3rd Baltic Conference on Human Language Technologies – the Baltic Perspective*. *Frontiers in Artificial Intelligence and Applications*. Čermák, František, Marcinkevičienė, Rūta, Rimkutė, Erika, Zabarskaitė, Jolanta (eds.). Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 235–240.
19. Paikens, Pēteris. 2016. Deep neural learning approaches for Latvian morphological tagging. *Human Language Technologies – The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016*. 289. *Frontiers in Artificial Intelligence and Applications*. Skadiņa, Inguna, Rozis, Roberts (eds.). Amsterdam: IOS Press, 136–143.
20. Paikens, Pēteris, Rituma, Laura, Pretkalniņa, Lauma. 2013. Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*. Oepfen, Stephan, Hagen, Kristin, Johannesen, Janne Bondi (eds.). Oslo: Linköping University Electronic Press, 267–277.
21. Rimkutė, Erika, Kovalevskaitė, Jolanta, Melninkaitė, Vida, Utkā, Andrius, Vitkutė-Adžgauskienė, Daiva. 2010. Corpus of contemporary Lithuanian language – the standardised way. *Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*. 219. *Frontiers in Artificial Intelligence and Applications*. Skadiņa, Inguna, Vasiljevs, Andrejs (eds.). Amsterdam: IOS Press, 154–160.
22. Rychlý, Pavel. 2007. Manatee/Bonito – a modular corpus manager. *RASLAN 2007 Proceedings (2007)*. Sojka, Petr, Horák, Aleš (eds.). Brno: Masaryk University, 65–70.
23. Skujiņa, Valentīna (red.). 2007. *Valodniecības pamatterminu skaidrojošā vārdnīca*. Rīga: LU Latviešu valodas institūts.

Summary

The main purpose of this paper is to present „The Balanced Corpus of Modern Latvian” (LVK) (www.korpus.lv) – a new 10 million representative corpus of contemporary Latvian. It describes the design, composition and text selection criteria of LVK2018. Also the annotation of the corpus (the metadata and the morphological tagging) and the usage of the corpus is described in the paper.

The history of the LVK series goes back to the 2007 when the first 1 million corpus was created. The LVK design, compilation and the text selection criteria were based on the

Latvian Language Corpus Conception. The same corpus design criteria were also used for the subsequent LVK series. The last corpus from that series (LVK2013) was released on 2013 with 4.5 million words. All corpora are morphologically annotated and the texts also annotated with metadata.

LVK2018 is the 10 million representative corpus of contemporary Latvian. LVK2018 is enlarged from LVK2013 based on the slightly modified corpus design criteria that also applied for the previous corpora from LVK series. LVK2018 is designed as general-language, representative and balanced corpus that aims to cover the variety of existing texts in some estimated proportions. The corpus contains five different sections: journalism (60%), fiction (20%), scientific (10%), legal (8%), parliamentary transcripts (2%).

This work has received financial support from European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219 (Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian).

Keywords: corpus; general corpus; balance; representativeness; metadata; text selection criteria.