Latvijas Universitāte
Matemātikas un informātikas institūts

LATVIEŠU VALODAS KORPUSA KONCEPCIJA

saīsināts variants
(netiek publicēta novecojusī informācija)

Pasūtītājs: Valsts valodas aģentūra

Rīga, 2005

Satura rādītājs

1. Valodas datu bāzu izveide Latvijā 4
1.1. Detalizēts pašreizējās situācijas izklāsts un analīze 4
1.2. Valodas resursu apstrāde Latvijā hronoloģiskā skatījumā 5
1.3. Pašreizējā situācija 6
1.4. Pašreizējās situācijas kopsavilkums 11
1.5. Latviešu valodas korpusa izveides nepieciešamības pamatojums 12
1.6. Korpusa izveides iespējamo problēmu un risinājumu raksturojums 12
1.7. Priekšlikumi korpusa izveidei Latvijā 14
2. Ārvalstu pieredzes apkopojums 23
2.1. Vispārīgie korpusi 23
2.2. Speciālie korpusi 33
2.3. Kopsavilkums 40
3. Sabiedriskā aptauja 44
4. Korpusa izmantojuma iespēju, pieejamības interesentiem un speciālistiem raksturojums 54
4.1. Vispārīgā latviešu valodas korpusa izmantošanas iespējas 54
4.2. Divvalodu (daudzvalodu) korpusu izmantošanas iespējas 66
4.3. Runas korpusa izmantošana 71
4.4. Valodas korpusa pieejamība 72
5. Latviešu valodas korpusa programmatūras izveides principu piedāvājums 76
5.1. Izstrādes vadlīnijas 76
5.2. Datu modelis 76
5.3. Programmrīki 82
5.4. Tekstu ievades principi 86
5.5. Kopsavilkums 86
6. Autortiesību (un autoratlīdzības) jautājums. Iespējamie risinājumi 89
6.1. Problēmas izklāsts (Ievads) 89
6.2. Lietotie termini 91
6.3. Tiesiskais regulējums atkarībā no korpusa lietojuma mērķa 91
6.4. Tekstu uzglabāšana un lietošana 93
6.5. Intelektuālo tiesību apjoms 94
6.6. Latviešu valodas korpusā iekļaujamo darbu veidi 95
6.7. Autora un citu autortiesību subjektu tiesības 100
6.8. Juridiskie nosacījumi darbu ievietošanai Latviešu valodas korpusā 106
6.9. Līguma būtiskās sastāvdaļas 117
6.10. Licence (vienkāršā, izņēmuma, vispārējā) 121
6.11. Mantisko tiesību kolektīvie pārvaldītāji 122
6.12. Juridiskie nosacījumi Latviešu valodas korpusa lietotājiem 124
6.13. Datu bāzes aizsargāšana 126
6.14. Atbildība par autortiesību pārkāpumiem 129
7. Sistēmas uzturēšana un paplašināšana 131
7.1. Korpusa sistēmas vispārējā arhitektūra 131
7.2. Programmatūras attīstība 134

8. Korpusa izstrādei nepieciešamā izpildes laika plānojums ar pamatojumu 136
8.1. Valodas korpusa izveides minimālā programma (5 gadi) 136
8.2. Valodas korpusa izveides maksimālā programma (+ 5 gadi) 138
8.3. Runas korpusa izstrāde 139
8.4. Paralēlā korpusa izveide 140
8.5. Laika plānojums 140
9. Latviešu valodas korpusa izveidei nepieciešamo izmaksu aprēķins. Citu ES valstu pieredze finansiālo jautājumu risināšanā. Iespējas izmantot ES fondu finansējumu 143
9.1. Latviešu valodas korpusa izveidei nepieciešamo izmaksu aprēķins, ņemot vērā piedāvātos risinājumus 143
9.2. Citu Eiropas Savienības valstu pieredze finansiālo jautājumu risināšanā 146
9.3. Iespējas izmantot Eiropas Savienības fondu finansējumu 148

1. Valodas datu bāzu izveide Latvijā

Detalizēts pašreizējās situācijas izklāsts un analīze, t. sk. latviešu valodas korpusa izveides nepieciešamības pamatojums un korpusa izveides iespējamo problēmu un risinājumu raksturojums, priekšlikumi korpusa izveidei Latvijā.

1.1. Detalizēts pašreizējās situācijas izklāsts un analīze

Pirms sākt latviešu valodas korpusa izveides koncepcijas izstrādi, ir jāaplūko pieejamie latviešu valodas resursi elektroniskā formā, kas koncepcijas 1. nodaļas virsrakstā saukti par valodas datu bāzēm. Ar **elektroniskajiem resursiem** koncepcijas autori saprot tekstu masīvu, kas tiek uzkrāts noteiktam mērķim.

Galvenie resursu veidi (to nosaukumi zināmā mērā pārklājas) ir:

- arhīvi;
- elektroniskās bibliotēkas;
- datorfonds;
- (runas, tekstu) korpuss.

Tekstu **arhīvi** ir tādu viegli lasāmu elektronisku tekstu krātuve, kuri nav nekādā veidā saskaņoti, piemēram, Oksfordas tekstu arhīvs [Atkins et al.1992].

Elektroniskās bibliotēkas ir elektronisku tekstu krājums standartizētā formātā ar īpašiem satura u. c. izveides noteikumiem, bet bez striktiem atlases kritērijiem [ibid].

Latvijā, pievēršot uzmanību latviešu valodas elektroniskajiem resursiem, sākumā tika runāts par latviešu valodas datorfonda veidošanu [Štekele 1994; Milčonoka 1995; Ozoliņa 1995; Spektors 1995] un par latviešu valodas datu bāzes izveidi [Kļaviņa 1993; Spektors, Baltiņa 1994]. Šobrīd Latvijā atrodam gan **elektroniskās bibliotēkas**[1], gan **datorfondu**[2], gan **tekstu korpusu**[3].

Korpuslingvistikā un datorlingvistikā ar **valodas korpusu** saprot rakstīta teksta vai transkribētas runas kopumu, ko izmanto valodas analīzei un aprakstam [sal. Kennedy 1998]. Tā kā terminam „korpuss" ir vairākas nozīmes, valodas apstrādes kontekstā tiek piedāvātas šādas korpusa definīcijas:

(a) jebkurš tekstu masīvs;

(b) parasti mašīnlasāms teksts;

(c) mašīnlasāmu tekstu izlase, kas veido maksimāli sabalansētu korpusu [McEnery, Wilson 2001].

Turpmāk tekstā, runājot par koncepciju, termins „valodas korpuss" tiks izmantots kā „mašīnlasāmu tekstu izlase, kas veido maksimāli sabalansētu korpusu".

[1] Folkloristikas elektroniskā bibliotēka, http://www.ailab.lv/FEB – skatīts 23.07.2005.
[2] Latviešu sakāmvārdu datorfonds, http://www.lfk.lv/sakamvardi – skatīts 23.07.2005.
[3] Latviešu valodas seno tekstu korpuss, http://www.ailab.lv/SENIE – skatīts 23.07.2005.

1.2. Valodas resursu apstrāde Latvijā hronoloģiskā skatījumā

Atskatoties uz datorlingvistikas vēsturi Latvijā, jāmin 60. gadu latviešu valodas funkcionālo stilu statistikas pētījumi: entropija un relatīvais burtu biežums latviešu alfabētā [Лоренц, Несауле 1963]; zinātniski tehnisko tekstu statistika [Якубайтис 1964]; prievārdu lietojums latviešu laikrakstos [Freidenfelds 1967]; publicistikas valodas biežuma vārdnīcas sastādīšana [Kļaviņa 1968]. Latviešu valodas institūtā tika uzsākts monumentāls pētījums – „Latviešu valodas biežuma vārdnīca" (1966 – 76), kurā sniegti funkcionālo stilu rādītāji, kas sakārtoti pēc alfabēta un biežuma.

Pirmais mašīntulkošanas eksperiments Latvijā no krievu valodas latviešu valodā, kas izmantoja latviešu un krievu valodas sakņu vārdnīcu un analīzes sintēzes likumus, tika veikts 60. gados [Гобземис и др. 1961].

Sevišķi ražīgi bija 70. gadi, kad, sadarbojoties Latviešu valodas un literatūras institūtam un Elektronikas un skaitļošanas tehnikas institūtam, tika uzsākts darbs ar automatizētu latviešu valodas morfoloģisko analīzi [Дризул 1972, 1974, 1976, 1978; Mecs 1970; Oša 1970]. Tika turpināts darbs pie datorizētas vārdnīcu sastādīšanas – iznāca biežuma un inversā biežuma vārdnīca. Latviešu valodas inversā vārdnīca iznāca 1970. g. [Soida, Kļaviņa 1970; papildinātais 2. izdevums 2000].

Liela uzmanība tika veltīta statistikas metodēm: tika pētīti valodas funkcionālie stili [Kļaviņa 1976], prievārdu lietojums [Zarovska 1977], grafēmu izplatība [Кузина 1977], tika sagatavots pārskats par latviešu valodas daiļliteratūras un zinātniski tehnisko tekstu grafēmu statistisko analīzi [Пиель 1988], kā arī pētījums par visu vārdšķiru statistiku [Якубайтис 1981]. Vienlaicīgi tika veikti eksperimenti par vārdšķiru varbūtisko saistāmību [Якубайтис, Скляревич 1978].

1966. – 76. g. tika publicēta pirmā latviešu valodas biežuma vārdnīca [Jakubaite et al. 1966 – 76], kas bija novatorisks mēģinājums izpētīt galvenos literārās valodas funkcionālos stilus, balstoties uz tekstu izlasēm. Par pamatu ņemot šo materiālu un veicot atlasi, M. Soikāne-Trapāne ASV 1985. g. publicēja „Latviešu valodas pamata un tematisko vārdu krājumu", kuru veido 1000 pamata vārdu un 19 tematiskas grupas (apm. 6000 vārdu). 1991. g. Rīgā iznāca „1000 vārdu. Latviešu valodas leksikas minimums ar tulkojumiem krievu un angļu valodā" [Bušs, Baldunčiks 1991]. Mūsdienu latviešu valodas sarunvalodas 3000 biežāk sastaptie vārdi, kas ir savākti laikā no 1976. – 1997. g., ierakstot sarunas veikalos, poliklīnikās u. c., tika publicēti 1998. g. [Kuzina 1998].

Folkloras materiālu uzkrāšana notika arī ārpus Latvijas. Te jāmin Bostonas Monreālas dainu korpuss[4] (K. Konrāde, V. Vīķe-Freiberga un I. Freibergs). Latvju dainu ievadīšanu datorā Latvijā 80. gadu beigās organizēja H. Bondars un S. Kļaviņa.

Astoņdesmitajos gados darbs pie statistikas pētījumiem tika sekmīgi turpināts Latviešu valodas institūtā [Пиель 1990] un LU Filoloģijas fakultātes Baltu valodu katedrā [Kļaviņa 1980]. Deviņdesmitajos gados LU Filoloģijas fakultātē tika turpināti latviešu valodas statistikas pētījumi, kā arī uzsākta 19. gs. dzejas tekstu uzkrāšana [Mūrniece 1997].

1.3. Pašreizējā situācija

1.3.1. *Iespiesto tekstu apstrāde*

Viens no lielākajiem latviešu valodas resursu turētājiem ir LU Matemātikas un informātikas institūta Mākslīgā intelekta laboratorija[5] (turpmāk tekstā – LU MII). Tekstu uzkrāšana elektroniskā formā tika aizsākta 80. gadu beigās, ievadot „Latviešu tautas ticējumu" fragmentus un lielākā 17. gs. avota – E. Glika tulkotās Bībeles – nodaļas [Spektors, Baltiņa 1994]. Tālāk tika turpināts darbs ar Sorosa fonds – Latvija un KKF atbalstu: tika uzkrāti gan folkloras[6], gan daiļliteratūras[7], gan citi materiāli.

Šobrīd LU MII nodarbojas ar diahronisko un sinhronisko resursu uzkrāšanu.

Diahroniskais aspekts

2002. gadā, sadarbojoties ar LU Filoloģijas fakultāti un Latvijas Nacionālo bibliotēku, agrāk uzkrātie senie teksti tika ievietoti Latviešu valodas seno tekstu korpusā[8], kas faktiski ir pirmais publiski pieejamais latviešu valodas tekstu korpuss ar programmrīkiem un dažādām, valodas korpusa analīzei piemērotām iespējām [Milčonoka 2003a].

Šobrīd seno tekstu korpusā pieejami 16. un 17. gs. svarīgākie iespiestie un rokrakstos esošie latviešu valodas teksti, kā arī nedaudzi 18. gs. teksti – galvenokārt garīga satura literatūra, daži lietišķie teksti un vārdnīcas. Pirms korpusa izveides daļa tekstu jau bija ievadīti manuāli, bet 17. gs. teksti tika ieskenēti ar datorprogrammu *ABBYY Fine Reader*, tādējādi paralēli arī tika izstrādāta šādu tekstus skenēšanas metodika. Pašlaik seno tekstu korpusā ir 30 avoti, kopīgais vārdlietojumu skaits pārsniedz 900 000 vārdlietojumus.

Seno tekstu korpusam tika izveidoti arī atbilstoši programmrīki, piemēram, tekstu pārbaudes un sagatavošanas programmrīki (specifiski seno tekstu korpusam), vārdformu

[4] http://www.president.lv/index.php?pid=011 – skatīts 21.07.2005.,
http://www.gramata21.lv/users/freibergs_imants – skatīts 21.07.2005.
[5] http://www.ailab.lv – skatīts 22.07.2005.
[6] http://www.ailab.lv/folklora – skatīts 22.07.2005.
[7] http://www.ailab.lv/Teksti – skatīts 22.07.2005.
[8] http://www.ailab.lv/SENIE – skatīts 22.07.2005.

indeksu veidošanas programmrīki, konkordances modulis u. c. Šī korpusa īpatnība ir iespēja aplūkot arī dažu korpusā ievietoto avotu faksimilattēlus.

Seno tekstu korpuss ir brīvi izmantojams zinātniskiem un pētniecības mērķiem. Tas tiek papildināts ar jauniem avotiem.

Sinhroniskais aspekts

No 1997. gada līdz 2000. gadam LU MII veica Latvijas Zinātnes padomes 11. nozares (valodniecība) granta 96.0245 „Latviešu valodas datorfonds" izpildi [Spektors 2001]. Pašlaik dažādu projektu ietvaros ir uzkrāts apmēram 25 miljonu liels elektronisko resursu masīvs [Levāne 2001], šie teksti galvenokārt tika uzkrāti, ieskenējot, kā arī nelielu daļu iegūstot elektroniskā formā.

90. gados LU MII sāka izstrādāt valodas korpusa marķēšanas metodiku un programmatūru (strukturālo, morfoloģisko un morfēmisko analizatoru). Tika izveidots strukturāls SGML marķēšanas rīks un uzkrāts apm. 225 000 vārdlietojumu liels marķēts, tekstu masīvs (sk. SGML marķētu tekstu dažādiem projektiem 1. pielikumā).

Ir veikts eksperimentāls projekts HTML marķētu daiļliteratūras darbu masīva automātiskai transformēšanai konsekventā, strukturālā XML formātā (sk. 2. pielikumā XML marķēta teksta paraugu).

Sadarbojoties ar B. Metuzāli-Kangeri [Metuzāle-Kangere 1985], tika strādāts pie morfēmiskā analizatora izveides [Sarkans 1995, 1998] (sk. 3. pielikumā morfēmiski analizēta teksta paraugu).

Paralēli tika arī attīstīta automatizēta morfoloģiskā analīze [Levāne, Spektors 2001; Greitāne 1997; Sarkans 1995; Spektors 2001]. Pašlaik morfoloģiski analizēts un manuāli pārbaudīts ir apmēram 15 000 vārdlietojumu liels tekstu masīvs. (sk. 4. pielikumā morfoloģiski marķētu tekstu ar manuālu pārbaudi un bez tās).

Jāsecina, ka LU MII ir uzkrāti plaši elektroniskie resursi un ir strādāts pie marķēšanas tehnoloģiju izveides, tomēr mūsdienīga valodas korpusa izveidei šie resursi būtu jāstrukturē, jāpapildina un būtu jāattīsta dažādu līmeņu marķēšanas sistēmas, veidojot sabalansētu un anotētu valodas korpusu.

1991. gadā tiek nodibināta sabiedrība „Tilde"[9], kas šobrīd izveidojusies par vadošo latviešu valodas tehnoloģiju izstrādes uzņēmumu. „Tildē" tiek veiksmīgi apvienota pētniecība un komercproduktu izstrāde: portālu indeksēšana; likumu korpusa uzkrāšana; morfoloģiskā analizatora izstrāde, kā arī tiek attīstīta mašīntulkošana. „Tilde" ar Valsts

[9] http://www.tilde.lv – skatīts 22.07.2005.

valodas aģentūras atbalstu ir izveidojusi pareizrakstības uzziņu sistēmu[10]. Sadarbībā ar LZA Terminoloģijas komisiju tapis terminoloģijas portāls[11].

1996. gadā tiek dibināts Tulkošanas terminoloģijas centrs (TTC), kas izveidots, lai „valsts pārvaldes iestādes un sabiedrību nodrošinātu ar valsts un starptautisko organizāciju izdoto tiesību aktu un citu dokumentu tulkojumiem, kā arī sniegtu priekšlikumus terminoloģijas izstrādes un standartizēšanas jomā"[12], šeit tiek uzkrāti dažādu veidu valodas resursi (piem., Eiropas Kopienas tiesu nolēmumi, kas šobrīd tiek tulkoti un kuru apjoms varētu sasniegt līdz 20 000 lappušu).

Latviešu Folkloras krātuvē tiek apkopoti dažādi folkloras materiāli, uzsākta arī tekstu digitalizēšana[13], piemēram, sākta Latviešu sakāmvārdu datorfonda veidošana[14]. Sadarbībā ar sabiedrību „Lursoft" izveidota Dainu skapja elektroniskā versija[15].

LZA Terminoloģijas komisijā dažādu projektu ietvaros tiek digitalizētas terminu vārdnīcas, veidojot elektroniskas terminu datu bāzes. Resursi ir atrodami terminoloģijas portālā[16].

LU Latviešu valodas institūtā tiek digitalizētas dažas vārdnīcas, bet pagaidām citi elektroniskie resursi veidoti netiek.

Jāmin arī Latvijas laikrakstu arhīvi un lielākais šādu resursu turētājs sabiedrība „Lursoft"[17], kuru krājumos ir praktiski visi Latvijā iznākošie laikraksti, tādējādi veidojot vienu no apjomīgākajiem preses korpusiem.

Jāsecina, ka Latvijā ir uzkrāti lieli elektroniskie resursi, tomēr:

(1) lielākoties tie nav gramatiski marķēti (ar dažiem izņēmumiem);

(2) teksti ir uzkrāti izklaidus un ir dažādu resursu turētāju īpašums.

Tādējādi latviešu valodas korpusa izveidotājiem, pirmkārt, būs jāpanāk vienošanās par jau esošo resursu izmantošanas iespēju; otrkārt, jāvienādo šo resursu formāts;

[10] http://www.letonika.lv/morphology – skatīts 21.07.2005.
[11] http://termini.letonika.lv/DesktopDefault.aspx – skatīts 21.07.2005.
[12] http://www.ttc.lv – skatīts 21.07.2005.
[13] http://www.lfk.lv/lasitava.html – skatīts 21.07.2005.
[14] http://www.lfk.lv/sakamvardi – skatīts 21.07.2005.
[15] http://www.lfk.lv/dainu_skapis.html – skatīts 21.07.2005.
[16] http://termini.letonika.lv/DesktopDefault.aspx – skatīts 23.07.2005.
[17] http://www.lursoft.lv – skatīts 22.07.2005.

treškārt, jāveic tekstu marķēšana, metodikas aprobēšana, izstrāde un piemērošana; ceturtkārt, jānodrošina valodas korpusa programmrīku izstrāde vai esošo rīku piemērošana.

1.3.2. *Runātā teksta apstrāde*

Ar runas pētniecību nodarbojas LU Filoloģijas fakultātē un LU MII.

LU Filoloģijas fakultātē 2000. gada martā tika dibināta Fonētikas un datorlingvistikas laboratorija (vadītājs J. Grigorjevs). Tajā „galvenais darbs tiek veltīts latviešu valodas skaņu sistēmas akustiskajai izpētei, veidojot pamatu akustiskam latviešu valodas skaņu sistēmas aprakstam, bez kā nav iespējama valodas tehnoloģijas attīstība Latvijā. Laboratorijā veiktie pētījumi ir nepieciešami mākslīgās runas sintēzei, automātiskai runas atšifrēšanai, kā arī runātāja identificēšanas sistēmu izstrādei"[18].

LU MII 1995. – 1997. g. projekta „ONOMASTICA-COPERNICUS" laikā IPA transkripcijā sagatavoja apmēram 250 000 latviešu valodā lietotos īpašvārdus (vārdus, uzvārdus, vietvārdus un uzņēmumu nosaukumus).

LU MII 2001. g. ar Latvijas Zinātņu padomes atbalstu ir aizsākts darbs pie runātās valodas korpusa izveides. Pašlaik ir apkopoti, digitalizēti un transkribēti (ne fonētiskajā transkripcijā) vairāki materiāli:

(1) multimediju mācīblīdzekļa „Ko tu teici?" izmantotās frāzes, vārdi un teikumi (ap 1300 vienību), ko ierunājuši 15 cilvēki (5 vīrieši, 7 sievietes un 3 bērni);

(2) materiāls, kas ieskaņots starptautiskā semināra „Valoda un tehnika 2000. Baltijas perspektīva" laikā 1994. gada 10. – 11. novembrī; tas ir sinhronais tulkojums no angļu valodas latviešu valodā, tulces ir divas sievietes; ilgums – aptuveni 8 stundas;

(3) speciāli sagatavots teksts „Rīga", kurā iekļautas visas iespējamās latviešu valodas fonēmas un fonēmu varianti, latviešu literārajā valodā vārdu sākumā iespējamie līdzskaņu savienojumi, kā arī atspoguļotas skaņu pārmaiņas; tekstu ierunājuši 50 cilvēki: 29 sievietes vecumā no 15 līdz 56 gadiem, 21 vīrietis vecumā no 15 līdz 63 gadiem.

LU Filozofijas un socioloģijas institūtā 1992. gadā tika uzsākts nacionālās mutvārdu vēstures projekts „Dzīvesstāsti"[19], tajā ir uzkrāts liels ierakstu krājums.

Runātās valodas dati atrodami arī LU Latviešu valodas institūtā, kur šobrīd notiek agrāk veikto dialektoloģijas ierakstu digitalizācija. Sabiedrības „Tilde" interešu sfērā ir

[18] http://www.lu.lv/filol/Baltu_nodala/Baltu_nod.htm – skatīts 25.07.2005.
[19] http://www.dzivesstasts.lv/lv/default.htm – skatīts 22.07.2005.

arī runas tehnoloģiju attīstība. Runātās valodas resursi ir Latvijas radiostaciju un TV sabiedrību fondos.

Apkopojot jāsecina, ka runātās valodas pētījumu un praktisko iestrāžu Latvijā nav daudz. Veidojot latviešu valodas runas korpusu, būtu jāapzina visi elektroniskā formā pieejamie runas materiāli.

1.3.3. *Divvalodu resursu uzkrāšana*

Divvalodu teksti (dokumenti plašā izpratnē, mājas lapas, rokasgrāmatas u. c.) tiek uzkrāti un apstrādāti daudzās LR valsts iestādēs, tulkošanas aģentūrās, pētniecības iestādēs un valsts un privātos uzņēmumos. Koncepcijas autori sīkāk minēs dažas iestrādes.

Tulkošanas un terminoloģijas centrā ir uzkrāti dokumenti, kurus varētu izmantot angļu un latviešu valodas paralēlajā korpusā (piem., Latvijas Republikas Civillikuma tulkojums angļu valodā; NATO dokumentācijas dokumenti; ES tiesību aktu tulkojumi latviešu valodā). Šobrīd kopīgais resursu apjoms sasniedzis 93 000 Oficiālā Vēstneša lappuses angļu valodā un 140 000 lappuses latviešu valodā.

LU MII 1995. – 2001. g. piedalījās ES projektā TELRI (*Trans – European Linguistic Resources Infrastructure*), šī projekta laikā uzsākts darbs pie paralēlo tekstu uzkrāšanas un apstrādes. Ir izveidots Platona „Valsts" paralēlo tekstu krājums ar 14 citām Eiropas valodām un izdots kompaktdisks ar šiem paralēlajiem tekstiem [Erjavec et al. 1998]. TELRI valodas rīku un resursu arhīvā ir pieejams teikuma līmenī sastatīts Dž. Orvela romāna „1984" tulkojums latviešu valodā ar oriģinālu. Abu tekstu sastatīšanai teikuma līmenī izmantots *Vanilla* alignators, kas izstrādāts Gēteborgas universitātē [Daniellson, Ridings 1997] un kas izmanto Čērča un Geila [Gale, Church 1993] algoritmu, teikumu sastatīšanā operējot ar simbolu skaitu.

ES likumdošanas teksti ir izmantoti mašīntulkošanas sistēmas izstrādei [Greitāne 1998]. Vienlaicīgi notiek pētījumi, kā informāciju, kas iegūta no paralēlajiem tekstiem, var izmantot MT sistēmās [Skadiņa 2001, 2002, 2003, 2005].

Pateicoties sadarbībai ar Tulkošanas un terminoloģijas centru, 2001. g. tika izveidots neliels (aptuveni 100 000 vārdu katrā valodā) eksperimentāls angļu-latviešu valodas korpuss, sastatot to teikuma līmenī. Tas izmantots angļu valodas darbības vārdu savienojumu un to tulkojuma ekvivalentu latviešu valodā analīzē [Milčonoka 2001a], kā arī tulkošanas pētījumos [Milčonoka 2001b, Milčonoka 2003b]. Izpētot paralēlā korpusa datus un vārdnīcās sniegto informāciju, secināts, ka viens no veidiem, kā papildināt vārdnīcu datus, ir atrast vienvalodu korpusā tipiskākos (biežāk sastopamos) šķirkļa

vārda vārdu savienojumus un meklēt to tulkošanas ekvivalentus paralēlajos tekstos [Milčonoka 2001b].

Internetā ir pieejams H. Celmiņas darbs „Sievietes padomju cietumos"[20] un tā tulkojums angļu valodā, bet nav veikta tekstu sastatīšana.

Plaši publiski pieejami daudzvalodu interneta resursi, kur pārstāvēta arī latviešu valoda, ir, piemēram, *Eur-Lex* tiesību aktu datu bāze[21]. Tiem nav veikta sastatīšana un citu veidu apstrāde.

Secinot par divvalodu tekstiem, jāmin, ka visbiežāk sastaptais valodu pāris ir latviešu-angļu valoda, bet ir pieejami arī latviešu-krievu valodas teksti. Taču divvalodu (daudzvalodu) tekstu pieejamība ir ļoti ierobežota, lielākoties (izņemot ES tiesību aktu dokumentus) tie publiski nav pieejami. Tāpēc valodas korpusa veidotājiem būs jāpanāk vienošanās ar lielākajiem resursu turētājiem un jāveic datu priekšapstrāde (piem., jāatlasa līdzīgi dokumenti, lai tie nedublētos), pirms tos varēs izmantot korpusa izveidē.

1.4. Pašreizējās situācijas kopsavilkums

1.4.1. *Tekstu korpuss*

Kā jau tika minēts, aplūkojot iepriekšējos gados un mūsdienās paveikto, daudzās iestādes ir uzkrāti dažādi elektroniskie resursi. Tāpat internetā pieejami daudzi teksti un datu bāzes, kaut gan daudzas datu bāzes ir slēgtas, piemēram, laikrakstu portāli. Pārsvarā šie elektroniskie resursi ir uzkrāti kādam konkrētam mērķim, tie nav strukturēti un gramatiski anotēti. Jāmin, ka LU MII, TTC un „Tildē" šiem resursiem daļēji ir veikta lingvistiska apstrāde, bet šie resursi praktiski nav pieejami pētniekiem un interesentiem. Ir nepieciešama plaša, sabalansēta korpusa izveide ar dažādiem programmrīkiem un iespējām dabīgās valodas pētniecībai, korpusam vajadzētu būt pieejamam publiski.

1.4.2. *Runas korpuss*

Runas korpusam nepieciešamie resursi ir uzkrāti daudz mazākā mērā nekā tekstu korpusam. Tas ir galvenokārt darīts tikai LU MII un LU Filoloģijas fakultātē. Turklāt uzkrātie resursi nav strukturāli marķēti vai arī strukturālais marķējums ir nepilnīgs. Jāmin, ka internetā ir atrodami plaši resursi (radio pārraides, arī dažādi televīzijas raidījumi), bet te ļoti būtiski ir panākt vienošanos, atrisinot autortiesību jautājumus, kā arī izstrādāt rīkus resursu automātiskai uzkrāšanai vēlamā formātā.

[20] http://www.ailab.lv/Teksti/Musdienas/Celmina/SievPSRS/saturs.htm – skatīts 22.07.2005.
[21] http://europa.eu.int/eur-lex/lex/lv/index.htm – skatīts 23.07.2005.

1.5. Latviešu valodas korpusa izveides nepieciešamības pamatojums

Lai mūsdienīgi aprakstītu un pētītu latviešu valodu dažādos līmeņos un nodrošinātu kvalitatīvu gramatiku, vārdnīcu, mācīblīdzekļu un dažādu mākslīgā intelekta sistēmu izstrādi, ir nepieciešams latviešu valodas korpuss, kas objektīvi reprezentē (modelē) valodu tās daudzveidībā.

Sakarā ar interneta straujo attīstību visā pasaulē valodas datorsistēmu attīstīšana mūsdienās kļūst par attiecīgās valodas izdzīvošanas jautājumu nākotnē [Spektors 2001].

Mūsdienās ir vairāki dabīgās valodas apstrādes virzieni, kuriem tiešā vai pastarpinātā veidā ir nepieciešams latviešu valodas korpuss, piemēram:

(1) vārdnīcu izstrāde;

(2) valodas pētniecība un modelēšana ar statistiskām metodēm;

(3) pareizrakstības un gramatikas automatizēta pārbaude;

(4) dialoga sistēmas programmrīki, kas lietotājam sniegtu iespēju brīvā latviešu valodā sazināties ar datu bāzēm un interaktīvām elektroniskajām sistēmām;

(5) tulkotāja programmrīki un mašīntulkošanas sistēmas;

(6) teksta semantiskās analīzes un informācijas izguves rīki;

(7) runas analīze un sintēze;

(8) multimediju valodas mācīblīdzekļu sagatavošana.

Valodas korpuss būs noderīgs arī daudzās citās nozarēs: pedagoģijā plašākā nozīmē, translatoloģijas pētījumos, literatūrzinātnē (autorības noteikšana, konkrēta autora vai daiļdarba stila valodas analīze) un sociolingvistikā.

Mūsdienīgu datorlingvistikas rīku izstrādei vispirms ir nepieciešams izveidot valodas resursus elektroniskā formā. Speciālisti uzskata, ka dabīgās valodas datoranalīzei vajadzīgs tekstu krājums elektroniskā formā, kas satur vismaz 150 miljonus vārdlietojumu. Jāpiebilst gan, ka šis vērtējums radies, balstoties uz angļu valodas pieredzi, un nav izslēgts, ka fleksīvām valodām būtu vajadzīgs vēl lielāks tekstu masīvs [Spektors 2001].

Korpusā balstītie rezultāti ietekmē gan teorētiskos pētījumus, gan veicina praktiskas izstrādes, tādējādi sniedzot ieguldījumu Latvijas tautsaimniecībā.

1.6. Korpusa izveides iespējamo problēmu un risinājumu raksturojums

Valodas korpusa izveide ir saistīta ar daudzām problēmām gan tehniskā, gan valodnieciskā, gan juridiskā ziņā. Šajā sadaļā problēmas un to risinājumi tikai tiks

ieskicēti, detalizētāka informācija ir atrodama turpmākajā valodas korpusa koncepcijas izklāstā.

1.6.1. *Korpusa pamatnostādnes*

Ir svarīgi vienoties par valodas korpusa izveides pamatnostādnēm: korpusa izmantošanas mērķi un tā saturu, korpusa tehnisko arhitektūru, programmnodrošinājumu un datu pieejamību un izmantošanu.

1.6.2. *Tekstu izvēle, ieguve un uzkrāšana*

Svarīgs ir datu kvalitātes un kvantitātes aspekts. Vai valodas korpuss būs noslēgts vai pastāvīgi papildināms? Kāda veida teksti būs valodas korpusā, kādas būs to proporcijas, kādā veidā tie tiks iegūti, kāds laika periods tiks aptverts? Kādā veidā tiks atlasīti teksti, lai tie būtu tipiski un reprezentatīvi? Vai korpusā tiks ievietoti teksti pilnībā vai tikai to fragmenti? Te jāņem vērā arī citu valstu pieredze šajos jautājumos. Reprezentatīva valodas korpusa izveidei ir nepieciešami dažāda laika perioda teksti, bet senāki teksti nav pieejami elektroniskā formātā, tāpēc tie ir jādigitalizē skenējot, kas tomēr ir laikietilpīgi, kā arī pēc tam ir nepieciešama ievadkļūdu novēršana (pieredze rāda, skenēšanas procesā ir sasniedzama precizitāte līdz 98%).

Elektronisku tekstu ieguve arī rada zināmas problēmas. Liela daļa tīmeklī atrodamie resursi nav izmantojami korpusā bez priekšapstrādes un bez vienošanās ar šo resursu turētājiem. Daļa resursu pieder izdevniecībām, par kuru izmantošanu arī jāvienojas. Iespējams, šajā gadījumā drīzāk varētu panākt vienošanos par tekstu fragmentiem, kas varētu būt pieņemams risinājums.

Lai korpusu periodiski papildinātu ar tīmeklī brīvi pieejamiem tekstiem (t. sk. ziņu portālu un forumu materiāliem), var izmantot t. s. „zirnekli".

Lai korpusa dati būtu mašīnlasāmi un saprotami, kā arī papildināmi un korpusa daļas savstarpēji savietojamas, svarīgi ir vienoties par uzkrājamo tekstu formātu – marķējuma standartiem (par to sīkāk sk. 5. nodaļu), kā arī tekstu dažādo versiju (piem., tīrs teksts, pārbaudīts teksts, anotēts teksts) uzturēšanu. Lai iegūtu viennozīmīgu marķējumu, jāizstrādā atbilstoša latviešu valodas korpusa marķēšanas metodika.

Lai atrisinātu gramatiskās un semantiskās neviennozīmības jautājumus, ir nepieciešamas teorētiskās nostādnes morfoloģijas, sintakses un citos jautājumos, kas latviešu valodniecībā nav līdz galam atrisināti vai vispār maz pētīti.

Lai runātās valodas dati būtu iekļauti sabalansētā korpusā, ir jānodrošina to ieguve no dažādiem avotiem (piem., vienojoties ar raidstacijām par radio translāciju ierakstīšanu, iegūstot konferenču u. tml. ierakstus). Arī runas korpusa daļas teksti ir jāuzkrāj atbilstošā formātā.

1.6.3. *Autortiesības*

Korpusa izveidošana un izmantošana ir cieši saistīta ar autortiesību jautājumu (par to sīkāk sk. koncepcijas 6. nodaļu).

1.6.4. *Valodas korpusa programmrīku izstrāde*

Programmrīku izstrādes efektivitāte un lietojumu funkcionalitāte ir tieši atkarīga no valodas korpusa marķējuma līmeņiem un to kvalitātes. Sk. 5. nodaļu par programmrīku izstrādi.

1.6.5. *Korpusa arhitektūra, administrēšana un uzturēšana*

Nav vienprātības par to, kādai jābūt korpusa sistēmas tehniskajai arhitektūrai. Iespējami divi galvenie virzieni, kas būtiski ietekmē administrēšanas un uzturēšanas jautājumus:

(1) centralizēta vai ierobežoti decentralizēta tīmekļa bāzēta sistēma;

(2) lokāla darbvirsmas sistēma (izplatāma, piem., CD formātā).

Par šo virzienu priekšrocībām un trūkumiem sīkāk skatīt 7. nodaļu par sistēmas uzturēšanu un paplašināšanu.

Lai nodrošinātu datu izmantošanu atbilstoši tam, kā datu turētāji ir atļāvuši tos izmantot (piem., tikai pētniecības un mācību mērķiem), ir jāizstrādā lietotāju piekļuves kontroles sistēma. Par to sīkāk skatīt 5. nodaļu un 7. nodaļu.

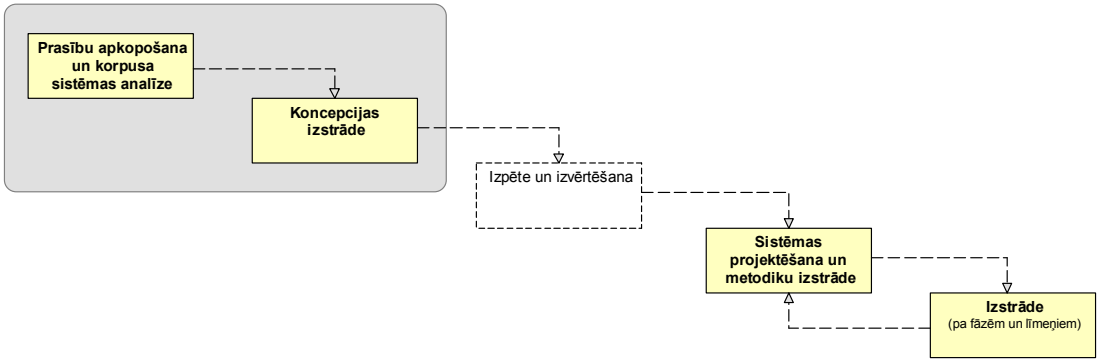
1.7. Priekšlikumi korpusa izveidei Latvijā

Koncepcijas autori piedāvā rekomendācijas pētniecības mērķiem publiski pieejamas latviešu valodas korpusa sistēmas izveidei, kas būtu piemērojama dažāda veida latviešu valodas korpusiem (vispārīgam, speciālam). Šādu sistēmu lielā mērā var attiecināt arī uz divvalodu korpusa izveidi, bet tomēr koncepcijas autori iesaka divvalodu korpusu uztvert kā atsevišķu vienību.

Šajā koncepcijā norādīti dažādi brīvi pieejami (arī valodneatkarīgi) programmlīdzekļi, no kuriem daļa, iespējams, ir piemērojama latviešu valodas korpusa vajadzībām. Dažādi esošie anotēšanas standarti, kuru piemērošana (apakškopu izvēle un iespējamā paplašināšana) latviešu valodai vēl ir detalizētāk jāizpētī un jāizvērtē.

Ņemot vērā izpētes rezultātus un šīs koncepcijas rekomendācijas, ir jāveic korpusa detalizēta projektējuma izstrāde.

Pabeidzot projektējuma izstrādi, var sākt veidot latviešu valodas korpusu. Paralēli projektēšanai var risināt datu izvēles un ieguves jautājumus.



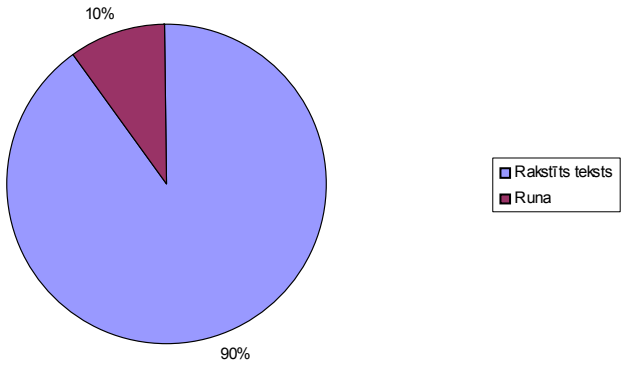
1.1. diagramma – korpusa izstrādes procesa „ūdenskrituma" modelis. Ar šo koncepcijas sagatavošanu ir aptverti pirmie divi posmi.

Koncepcijas autori iesaka izveidot vairāku tipu latviešu valodas korpusus: vispārīgo, speciālo un divvalodu (vai daudzvalodu) korpusu. Sīkāk tiks apskatīta katra veida struktūra.

1.7.1. *Latviešu valodas vispārīgais korpuss*

Valodas korpusa koncepcijas izstrādes laikā tika meklēti risinājumi valodas korpusa izveidei 5–10 gadu perspektīvā. Tādējādi darbam ir paredzēti vairāki posmi. Šeit gan jāmin, ka darba grafiks ir atkarīgs no piešķirtā finansējuma un cilvēkresursiem.

Balstoties uz pasaules pieredzi nacionālo korpusu veidošanā, koncepcijas autori piedāvā sākumā izveidot 1 milj. vārdlietojumu mūsdienu (sākot ar 20. gs. 80. gadiem) latviešu valodas korpusu, ievērojot šādu mutvārdu un rakstveida tekstu proporciju:



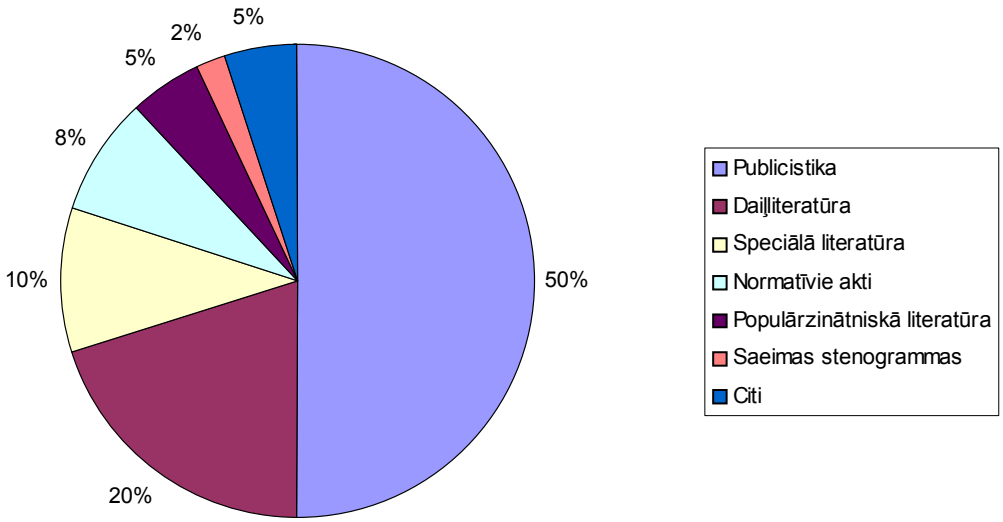
1.2. diagramma – runātās un rakstītās valodas attiecība vispārīgajā latviešu valodas korpusā.

Lai izveidotu sabalansētu un reprezentatīvu latviešu valodas korpusu, koncepcijas autori iesaka ietvert dažādu funkcionālo stilu tekstus. Vēlams, lai korpusā būtu pārstāvēti dažādu rakstīto avotu dati:

– publicistika (laikraksti, žurnāli, interneta publicistikas materiāli piemēram, ziņas no dažādiem portāliem) – 30 – 50%;

– speciālā literatūra (zinātniski teksti) – 10 – 20%;
– daiļliteratūra (oriģinālliteratūra, tulkotā literatūra) – 10 – 30%;
– normatīvie akti (likumi, regulas u. c.) – 5 – 10%;
– populārzinātniskā literatūra – 5 – 10%;
– Saeimas stenogrammas – 2 – 5%;
– citi teksti – 5 – 10%.

Viens no piedāvātajiem modeļiem būtu:



1.3. diagramma – rakstītās valodas attiecības korpusā.

Attiecībā uz runātās valodas komponenti koncepcijas autori iesaka iekļaut dialogus un monologus vienādās attiecībās. Koncepcijas autori piedāvā šādu sīkāku monologu sadalījumu:

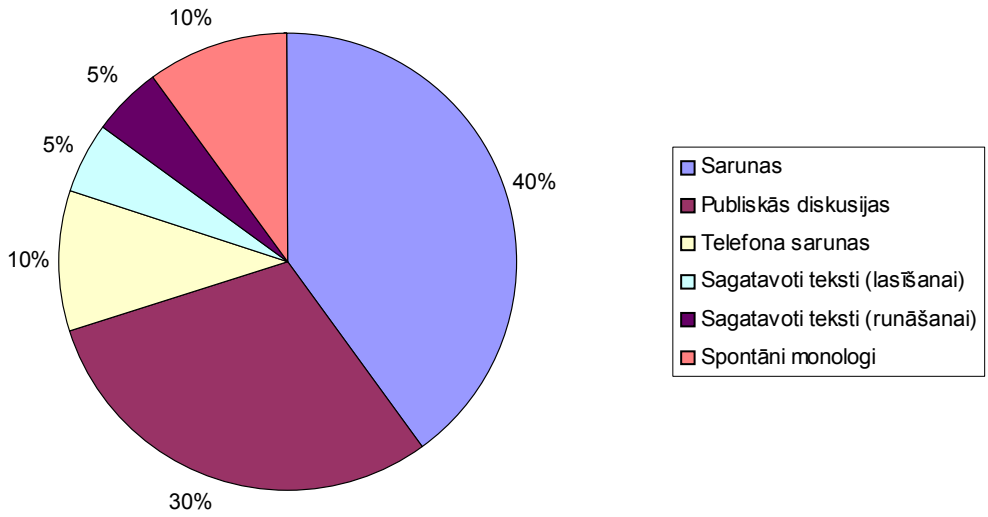
(1) lasīšanai iepriekš sagatavoti teksti – visu veidu ziņas (sporta, laika, ekonomikas u. tml.);
(2) runāšanai sagatavoti teksti – lekcijas, spredi̇ķi, parlamentāriešu runas u. c., kas netiek lasīti, bet runāti pēc sagatavota plāna, tēzēm;
(3) spontāni monologi – sporta sacensību komentāri, stāstījums (piemēram, cilvēku atmiņu stāstījums).

Savukārt dialogiem paredzēts šāds dalījums:

(1) telefonu sarunas;
(2) sarunas, cilvēkiem atrodoties vienā telpā;

(3) publiskā diskusija (piem., J. Dombura raidījums „Kas notiek Latvijā?", radio un televīzijas intervijas).

Viens no piedāvātajiem modeļiem būtu:



1.4. diagramma – runātās valodas attiecības korpusā.

Korpusa projekta izstrādes laikā viss korpuss ir atbilstoši jāmarķē (gan strukturāli, gan gramatiski) un jānodrošina attiecīgo programmrīku izstrāde (vai esošo rīku piemērošana latviešu valodas korpusa vajadzībām). Papildus marķēšanas sistēmas izstrādei ir jāparedz arī korpusa datu metainformācijas modeļa izveide un tā īstenošana, kas uzlabo tekstu atlasi (ļauj galalietotājam precizēt korpusa apstrādes apgabalu).

Lai nodrošinātu korpusa funkcionālu izmantošanu, ir jārealizē izvēlētā korpusa arhitektūra (saskarne) ar programmrīkiem (piem., konkordances lietojums).

Korpusa izveides laika plānojumu un darbu aprakstu skatīt 8. nodaļā.

Koncepcijas autori piedāvā izveidot arī **latviešu valodas speciālos korpusus**. Šobrīd tiek ieteikta **izlokšņu korpusa** un **studentu (valodas apguvēju) korpusa** izveide.

Ņemot vērā, ka Latvijā plaši tiek veikti dialektoloģijas pētījumi, ir vērts dažādās iestādēs sakrātos materiālus apvienot speciālā korpusā. Līdz šim ekspedīcijas materiāli ir uzkrāti audio kasetēs un to atšifrējumos. Lielākās problēmas ir saistītas ar šāda materiāla atšifrējuma vienādošanu, jo, kā zināms, eksistē dažādi pierakstu veidi.

Vajadzētu izveidot savu viennozīmīgu standartu, papildinot pašlaik izmantoto mašīnlasāmo fonētisko alfabētu. Turklāt izveidotajam standartam jābūt atgriezeniski savietojamam ar vispārpieņemtiem standartiem. Domājams, ka izlokšņu korpusā vajadzētu iekļaut:

(1) izloksnes teksta pierakstu ortogrāfijā;

(2) mašīnlasāmu fonētiskās transkripcijas pierakstu;

(3) „tulkojumu" literārajā valodā.

1.7.2. *Studentu (valodas apguvēju) korpuss*

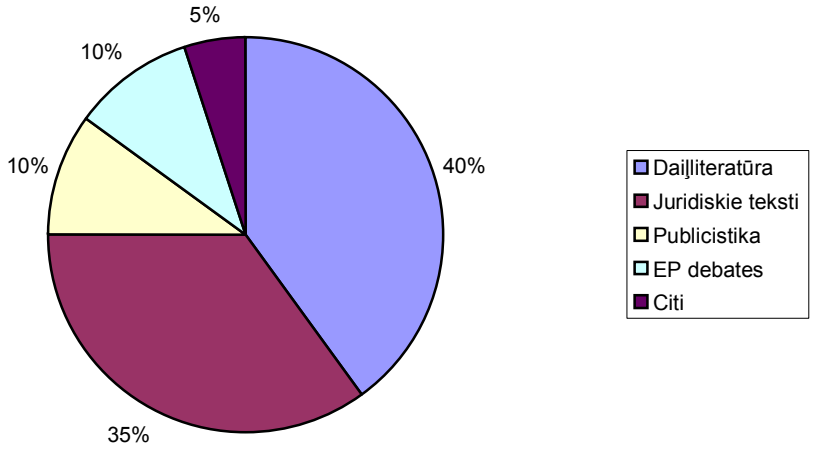
Ņemot vērā, ka Latvijā latviešu valodu apgūst arī daudzi cittautieši, nozīmīga ir latviešu valodas studentu (valodas apguvēju) korpusa izveide. Turklāt tādā korpusā var pievienot arī to ārzemju augstskolu studentu datus, kas mācās latviešu valodu. Parasti šādu mācību korpusu veido studentu (valodas apguvēju) sacerējumi par noteiktām tēmām. Šeit īpaši tiek marķētas valodas kļūdas. Tādu korpusu izmanto kļūdu analīzē un uz piemēriem balstītā valodas apguvē, kā arī tas veicina jaunu mācīblīdzekļu un gramatiku izstrādi.

1.7.3. *Divvalodu vai daudzvalodu korpuss*

Ņemot vērā mūsdienu situāciju, koncepcijas autori piedāvā izveidot arī **divvalodu (vai daudzvalodu) korpusu**. Pasaules prakse liecina, ka divvalodu (daudzvalodu) korpusus visbiežāk sauc par **paralēlajiem korpusiem**, speciāli nenošķirot, vai tas ir tulkojumu vai salīdzināmais korpuss. Domājot par nepieciešamajiem valodu pāriem, koncepcijas autori secinājuši, ka ir nepieciešamība pēc latviešu-angļu, latviešu-krievu un arī latviešu-franču valodas paralēlā korpusa. Koncepcijas autori iesaka sākt ar latviešu-angļu valodas paralēlo korpusu.

Divvalodu korpusi mēdz būt vienvirziena (*mono-directional*), resp., tajos ir tikai vienas valodas tulkojumi otrā, vai divvirziena (*bi-directional*), kad korpusā ir tulkojumi abās valodās. Pētniecībai būtu vēlams izveidot divvirziena korpusu, kurā būtu pārstāvēti gan latviešu valodas oriģināldarbi, gan tulkojumi latviešu valodā.

Vēlams, lai paralēlajā korpusā būtu pārstāvēti dažādu funkcionālo stilu teksti:



1.5. diagramma – paralēlā korpusa tekstu sadalījums.

Viens no lielākajiem divvalodu resursu turētājiem ir Tulkošanas un terminoloģijas centrs, tāpēc koncepcijas autori ir tikušies ar tā direktoru M. Baltiņu (05.07.2005.), lai apspriestu iespējas izmantot viņu rīcībā esošos juridiskos tekstus un to tulkojumus korpusa izveidē. Tikšanās laikā Tulkošanas un terminoloģijas centrs ieteica noslēgt sadarbības līgumu par noteikta daudzuma angļu un latviešu dokumentu atlasi (tādējādi novēršot saturā līdzīgu dokumentu dublēšanu) korpusa vajadzībām.

Šie ir konkrēti priekšlikumi par iespējamo piedāvāto valodas korpusu saturu. Sīkāk par programmnodrošinājumu u. c. jautājumiem skatīt pārējās koncepcijas nodaļās.

Vēres

Atkins S., Clear ., Ostler J. [1992], "Corpus design criteria." – *Literary and Linguistic Computing* 7(1). – pp. 1–16.

Auziņa I. [2004], „Latviešu valodas grafēmas-fonēmas atbilsmju likumu sistēma." – *Latvijas Zinātņu Akadēmijas Vēstis* 58(3). – Rīga – 11.–18. lpp.

Bušs O., Baldunčiks J. [1991], „1000 vārdu**:** Latviešu valodas leksikas minimums ar tulkojumu krievu un angļu valodā." – *LZA VLI*, Atb. red. O. Bušs. – Rīga: Zinātne, 1991. – 48. lpp.

Danielsson P., Ridings D. [1997], "Practical Presentation of a "Vanilla" Aligner." Presentation held at the TELRI Workshop in alignment and exploitation of texts in Ljubljana, Feb. 1–2. – http://nl.ijs.si/telri/Vanilla/doc/ljubljana – skatīts 25.07.2005.

Erjavec T., Lawson A., Romary L. [1998], "East meets West: Producing Multilingual Resources in a European Context." – *First International Language Resources and Evaluation Conference*. – Granada, Spain.

Freidenfelds I. [1967], „Prievārdu lietošanas biežums latviešu laikrakstos." – *Leksikoloģijas un leksikogrāfijas jautājumi*. – Rīga: LVU – 40.–47. lpp.

Gale W.A., Church K.W. [1993], "A Program for Aligning Sentences in Bilingual Corpora. " – *Computational Linguistics*, Volume 19, Number 1 – pp. 75–90.

Greitāne I. [1997], „Mašīntulkošanas sistēma LATRA.” – *LZA Vēstis*, Nr.3/4 – 1–6. lpp.
Greitāne I., [1998], “Machine Translation and Multilingual Resources for Latvian.” – *Proceedings of the Third European Seminar “Translation Equivalence”*. – pp. 79.–86.
Grūzītis N., Auziņa I., Bērziņa-Reinsone S., Levāne-Petrova K., Milčonoka E., Nešpore G., Spektors A. [2004], “Demonstration of resources and applications at the Artificial Intelligence Laboratory,” IMCS, UL. – *Proceedings of the first Baltic conference "Human Language Technologies – the Baltic Perspective.* – Riga – pp. 38–42.
Jakubaite T., Kristovska D., Ozola V., Prūse R., Sika N. [1966.–1967], *Latviešu valodas biežuma vārdnīca.* – Rīga: Zinātne.
1. sēj. 1. d.: *Tehnika un rūpniecība*. Atb. red. T. Jakubaite. – 1966. – 624. 1pp.;
1. sēj. 2. d: *l. daļā. ietverto gramatisko kategoriju statistika.* – 1968. – 127. lpp.;
2. sēj. 1. d.: *Laikraksti un žurnāli.* – 1969. – 857. lpp.;
2. sēj. 2. d.: *2. sēj. 1. daļā. ietverto gramatisko kategoriju statistika.* – 1969. – 185. lpp. Autori abām daļām T. Jakubaite, D. Guļevska, V. Ozola, A. Rubīna, N. Sika;
3. sēj. 1. d.: *Daiļliteratūra*. – 1972. – 1154. lpp. Autori T. Jakubaite, D. Guļevska, V. Ozola, A. Rubīna, N. Sika;
Apvienotais (1–3) sēj. – 1973. – 1004. lpp. Atb. red. T. Jakubaite;
4. sēj.: *Zinātne.* – 1976. – 646. lpp. Atb. red. T. Jakubaite.
Kennedy G. [1998], *An Introduction to Corpus Linguistics*. – London: Longman.
Kļaviņa S. [1968], „Latviešu publicistikas valodas biežuma vārdnīca.” – *P. Stučkas Latvijas Valsts universitātes Zinātniskie raksti* 86. sēj. – Rīga. – 199.–216. lpp.
Kļaviņa S. [1976], „Korelācija vārdu krājumā un vārdšķiru lietojumā starp dažādu funkcionālo stilu tekstiem.” – *Latvijas PSR Zinātņu Akadēmijas Vēstis* Nr. 2. – Rīga – 112.-122. lpp.
Kļaviņa S. [1980], *Statistika valodniecībā.* – Rīga: LVU – 145. lpp.
Kļaviņa S. [1993], „Mūsdienu latviešu valodas vārdu krājuma datu bāze.” – *Baltu filoloģija* III. – 76.–78. lpp.
Kuzina V. [1998], *3000 latviešu sarunvalodas biežāk lietotie vārdi ar tulkojumiem krievu, vācu un angļu valodā.* – Rīga.
Levāne K. [2001], „Latviešu valodas korpuss, tā anotēšana un analīze.” – *Материалы межвузовской научно-методической конференции студентов.* – Санкт-Петербург – 34. lpp.
Levāne K., Spektors A. [2000], “Morphemic Analysis and Morphological Tagging of Latvian Corpus.” – *Proceedings LREC 2000.* – Athens – pp. 1095–1098.
McEnery T., Wilson A. [2001], *Corpus Linguistics*, 2nd Edition. – Edinburg University Press.
Mecs N. [1970], „Latviešu valodas morfoloģiskās analīzes algoritms un tā sastādīšanas algoritmizācijas problēmas.” – *Latviešu valodas struktūras jautājumi*. – Rīga – 241.–285. lpp.
Metuzāle-Kangere B. [1985], *Latviešu valodas atvasinājumu vārdnīca.*, – Hamburg: H. Buske Verlag – 392. lpp.
Milčonoka E. [1995], *Eduarda Veidenbauma prozas datorfonds.* Bakalaura darbs. – Rīga: Latvijas Universitāte.
Milčonoka E. [2001a], “The Contrastive Study of English Multi-word Verb Units and Their Translation Equivalents Based on English-Latvian Parallel Corpus.” *TELRI Newsletter,* 12. – pp. 17–18.

Milčonoka E. [2001b], "Some observations about English-Latvian Translation Equivalents in a new Bible for Europe: A Study Based on the EU legislation and its translation." – *COMPLEX2001, 6th Conference on Computational Lexicography and Corpus Research*. – Birmingham – pp. 175–187.

Milčonoka E. [2003a], „Latviešu valodas 17. gadsimta teksti internetā." – *Baltu filoloģija* XII (1) – Rīga – 139.–150. lpp.

Milčonoka E. [2003b], "Use of parallel corpora in translation studies." – *Terminology and Technology Transfer in the Multilingual Information Society*. *Proceedings of the 2nd International Conference on Terminology*. – TermNet Publisher & LLI of LU – pp. 78–98.

Mūrniece B. [1997], „Lingvostatistika autora valodas pētījumos." *Konferences "Valodas pētīšanas metodes" tēzes*. – Rīga: LU – 13. lpp.

Oša M. [1970], „Latviešu valodas morfoloģiskās analīzes algoritma pārbaude elektronu skaitļojamā mašīnā BESM-2." – *Latviešu valodas struktūras jautājumi.* – Rīga – 285.–289. lpp.

Ozoliņa A. [1995], „17. gs. tekstu datorfonda izveides programmlīdzekļi." – *Baltistica VII Starptautiskais baltistu kongress*. – Rīga – 83. lpp.

Sarkans U. [1995], "Morphemic and Morphological Analysis of the Latvian Language." – *Papers in Computational Lexicography COMPLEX'96.* – Budapest – pp. 219–226.

Sarkans U. [1998], „Mašīnmācīšanas metožu izmantošana latviešu valodas morfēmiskās analīzes projektā." – *Baltu filoloģija VIII* – Rīga: LU – 43.–47. lpp.

Skadiņa I. [2001], "Studies of English-Latvian Legal texts for Machine Translation." – *Abstract of paper at the 6th Telri Seminar*, Bansko, 8-11 November 2001, Telri Newsletter 12, October 2001, – pp. 13–14.

Skadiņa I. [2002], „Datortehnoloģijas lietojums tulkošanā un vārdnīcu izstrādē." – *E. Drezena (1892-1937) piemiņai veltītā 2. starptautiskā terminoloģijas konferences „Terminoloģija un tulkošanas tehnoloģija daudzvalodu informācijas sabiedrībā" referātu tēzes.* – Rīga – 44. lpp.

Skadiņa I. [2003], "Electronic Dictionaries and Multilingual Information Society." – *Terminology and Technology Transfer in the Multilingual Information Society.* – Termnet Publisher – pp. 140–146.

Skadiņa I. [2005], "Studies of English-Latvian Legal texts for Machine Translation." – *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora,* Continuum. – London, New York – pp. 188–195.

Soida E., Kļaviņa S. [1970], *Latviešu valodas inversā vārdnīca.* – Rīga.

Soida E., Kļaviņa S. [2000], *Latviešu valodas inversā vārdnīca.* – 2. papild. un lab. izd. – Rīga: RaKa – 396. lpp.

Soikāne-Trapāne M. [1985], *Latviešu valodas pamata un tematisks vārdu krājums. (Latvian basic and topical vocabulary)*. – Amerikas latviešu apvienība.

Spektors A., Baltiņa M. [1994], „Latviešu valodas vēsturisko tekstu datu bāzes izveide." – *Valoda un tehnika Eiropā 2000.* – Rīga – 30. lpp.

Spektors A. [1995], „Latviešu valodas datorfonds." *Baltistica VII*, Starptautiskais baltistu kongress – Rīga – 105.–106. lpp.

Spektors A. [1998], „Latviešu valodas datorlingvistikas resursi." – *Baltu filoloģija* VIII. – Rīga: LU – 53–59. lpp.

Spektors A. [2001], „Latviešu valodas datorfonda izveide." – *LZA Vēstis A* Nr. 2, – 74–82. lpp.

Štekele B. [1994], *19. gadsimta 90. gadu latviešu dzejas valodas datorfonds.* Maģistra darbs. – Rīga: Latvijas Universitāte.

Zarovska R. [1977], „Prievārdu lietojuma statistika.” – *Статистика и функциональные стили язика.* – Rīga – 77.–87. lpp.

Гобземис А. Ю., Горобец В.Г., Юрик В.А., Якубайтис Т.А. [1961], “О машинном переводе с русского языка на латышский.” – *Автоматика и вычеслительная техника*, №9. – стр. 149–164.

Дризул В. А. [1972], “Автоматический морфологический анализ текстов латышского языка.” – *Latvijas PSR Zinātņu Akadēmijas Vēstis* Nr. 6. – Rīga – 74.–82. lpp.

Дризул В. А. [1974], “Автоматическая обработка лингвистических данных методом аналогии.” – *Latvijas PSR Zinātņu Akadēmijas Vēstis* Nr. 2. – Rīga – 90.–93.lpp.

Дризул В. А. [1976], *Использование методов математической лингвистики для изучения латышского языка.* – Рига: Зинатне.

Дризул В. А. [1978], “Об автоматическом распознавании омонимии флексий латышского языка.” – *Latvijas PSR Zinātņu Akadēmijas Vēstis* Nr. 10. – Rīga 79.–87. lpp.

Кузина В. [1977], “Статистика букв в текстах разных типов современного латышского языка.” – *Latvijas PSR Zinātņu Akadēmijas Vēstis*. – Rīga – 97.–106. lpp.

Лоренц А., Несауле З. [1963], “Статические свойства латышского языка.” *Latvijas PSR Zinātņu Akadēmijas Vēstis* Nr. 10. – Rīga – 41.–48. lpp.

Милчонока Э. [2002], “Корпусная лингвистика и историческая лексикография.” – *Материалы XXXI межвузовской научно-методической конференции преподавателей и аспирантов.* Выпуск 1. Секция балтистики. Тезисы докладов. – Санкт-Петербург – стр. 34.

Милчонока Э. [2002a], “Обзор ресурсов латышского языка в Институте математики и информатики Латвийского университета.” – *Доклады научной конференции “Корпусная лингвистика и лингвистические базы данных”,* под ред. А. С. Герда. – Изд-во С.-Петерб. ун-та – стр. 108–123.

Милчонока Э. [2003b], “На пути к историческому словарю латышского языка.” – *Теоретическая лексикография: современные тенденции развития*, Материалы V международной школы-семинара. – Иваново – стр.193–194.

Милчонока Э. [2004], “Корпус старолатышских текстов.” – *Международная конференция “Корпусная лингвистика – 2004”: Тезисы докладов.* – Санкт-Петербург: Издательство СПбГУ – стр. 60–61.

Пиель Е. Ш. [1988], “Статистика графем латышских научно технических и художественных текстов.” – *Latvijas PSR Zinātņu Akadēmijas Vēstis* Nr. 11. – Rīga – 73.–83. lpp.

Пиель Е. Ш. [1990], “Текстовая база данных как эмпирическая база компьютерной лексикографии.” *Актуальные проблемы компьютерной лингвистики*. – Тарту – стр. 117.

Якубайтис Т. А. [1964], “О статической структуре научно технических текстов латышского языка.” – *Latvijas PSR Zinātņu Akadēmijas Vēstis* Nr. 3. – Rīga – 21.–25.lpp.

Якубайтис Т. А. [1981], *Части речи и типы текстов.* – Рига.

Якубайтис Т. А., Скляревич А. Н. [1978], *Вероятностные характеристики связных текстов.* – Рига.

2. Ārvalstu pieredzes apkopojums

Šajā apskatā sniegta informācija par dažādu tipu korpusiem. Ne visa korpusa informācija ir publiski pieejama (jo īpaši detalizēts korpusa marķējuma raksturojums), tāpēc jāņem vērā, ka koncepcijas autori sniedz tikai galveno informāciju par katru korpusu.

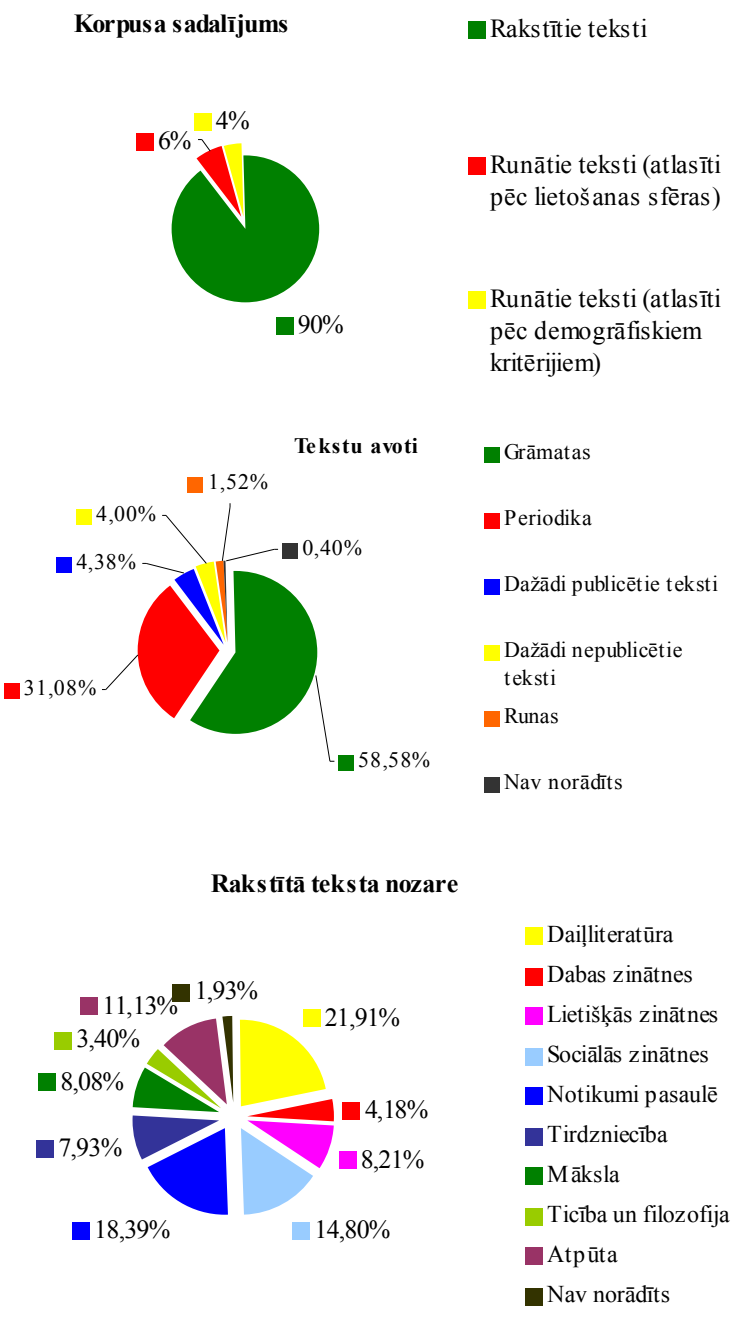
Šajā nodaļā sniegta informācija par šādiem korpusiem:

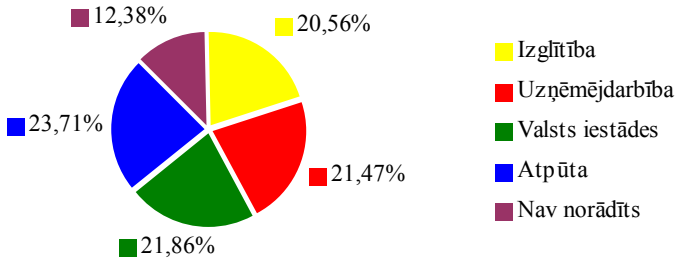
1) Britu nacionālo korpusu,
2) Čehu nacionālo korpusu,
3) Grieķu valodas korpusu (Hellēņu nacionālo korpusu),
4) Horvātu nacionālo korpusu,
5) Īru valodas nacionālo korpusu,
6) Krievu valodas nacionālo korpusu,
7) Lietuviešu valodas tekstu korpusu,
8) Poļu valodas nacionālo korpusu,
9) Zviedru runātās valodas korpusu,
10) Norvēģu-angļu paralēlo korpusu,
11) Oslo daudzvalodu korpusu,
12) Bergenas Londonas pusaudžu valodas korpusu (COLT),
13) Starptautisko angļu valodas mācību korpusu (ICLE),
14) 22 valodu telefona sarunu korpusu,
15) Somijas zviedru valodas tekstu korpusu,
16) Igauņu valodas dialektu korpusu,
17) Dikensa korpusu,
18) Sintaktiski anotētu vācu laikrakstu korpusu.

2.1. Vispārīgie korpusi

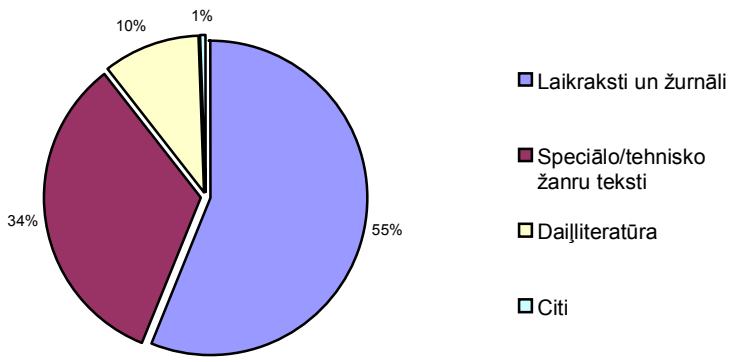
2.1.1. *Tekstu korpusi*

Nosaukums	**Britu nacionālais korpuss**
Tīmekļa vietne	http://www.natcorp.ox.ac.uk
Tips	Sinhronisks vispārīgs vienvalodas runas un rakstīto tekstu korpuss
Apjoms	100 milj. vārdlietojumu; teksti – 90%, runa – 10%. Korpuss ir pabeigts un netiek papildināts.

Autortiesības	Ar tekstu devējiem tiek noslēgta standarta vienošanās par tekstu izmantošanu pētniecības nolūkiem. Neviens teksts netiek iekļauts pilnībā, tiek izmantoti tikai fragmenti. Gala lietotājiem ar korpusa veidotājiem jānoslēdz līgums, lai par maksu varētu izmantot korpusu (tiešsaistē vai ar CD).
Finansējums	Zinātnes un tehnikas padome, Tirdzniecības un rūpniecības departaments, Britu bibliotēka un Britu akadēmija
Sadalījums	

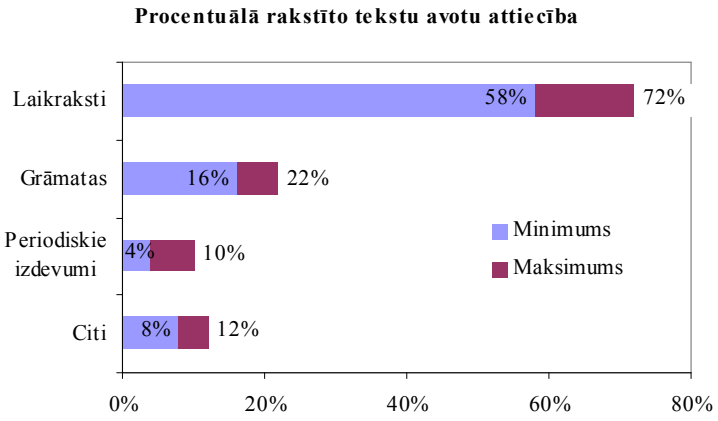
<table>
<tr><td></td><td colspan="3">Runātā teksta lietošanas sfēra
 
</td></tr>
<tr><td rowspan="6">Marķējums</td><td colspan="2">Līmeņi</td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="5">TEI vadlīnijas
SGML formāts</td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td>Morfoloģija</td><td>X</td></tr>
<tr><td>Sintakse</td><td>X</td></tr>
<tr><td>Fonētiskā transkripcija</td><td>X</td></tr>
<tr><td rowspan="2">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>CLAWS stochastic part-of-speech tagger[22]</td><td colspan="2">Segmentācija un vārdšķiru automatizēta marķēšana</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Čehu nacionālais korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://ucnk.ff.cuni.cz/english</td></tr>
<tr><td>Tips</td><td colspan="3">Sinhroniskais korpuss:
– sabalansēts mūsdienu tekstu korpuss;
– runas korpuss;
– ieplānots dialektu korpuss.
Diahroniskais korpuss.
Papildus:
– mašīnlasāmas vārdnīcas un citas datu bāzes;
– plašs neapstrādātu, elektronisku tekstu masīvs.</td></tr>
<tr><td>Apjoms</td><td colspan="3">Sinhroniskais korpuss: teksti – 100 milj. vārdlietojumu, runa – 700 tūkst. vārdlietojumu.
Diahroniskais korpuss: 2,3 milj. vārdlietojumu.</td></tr>
<tr><td>Izmantošana</td><td colspan="3">Korpusa izveides galvenais mērķis ir kalpot par pamatu jaunām kvalitatīvām vārdnīcām. Korpuss paredzēts arī plašiem valodas pētījumiem.</td></tr>
</table>

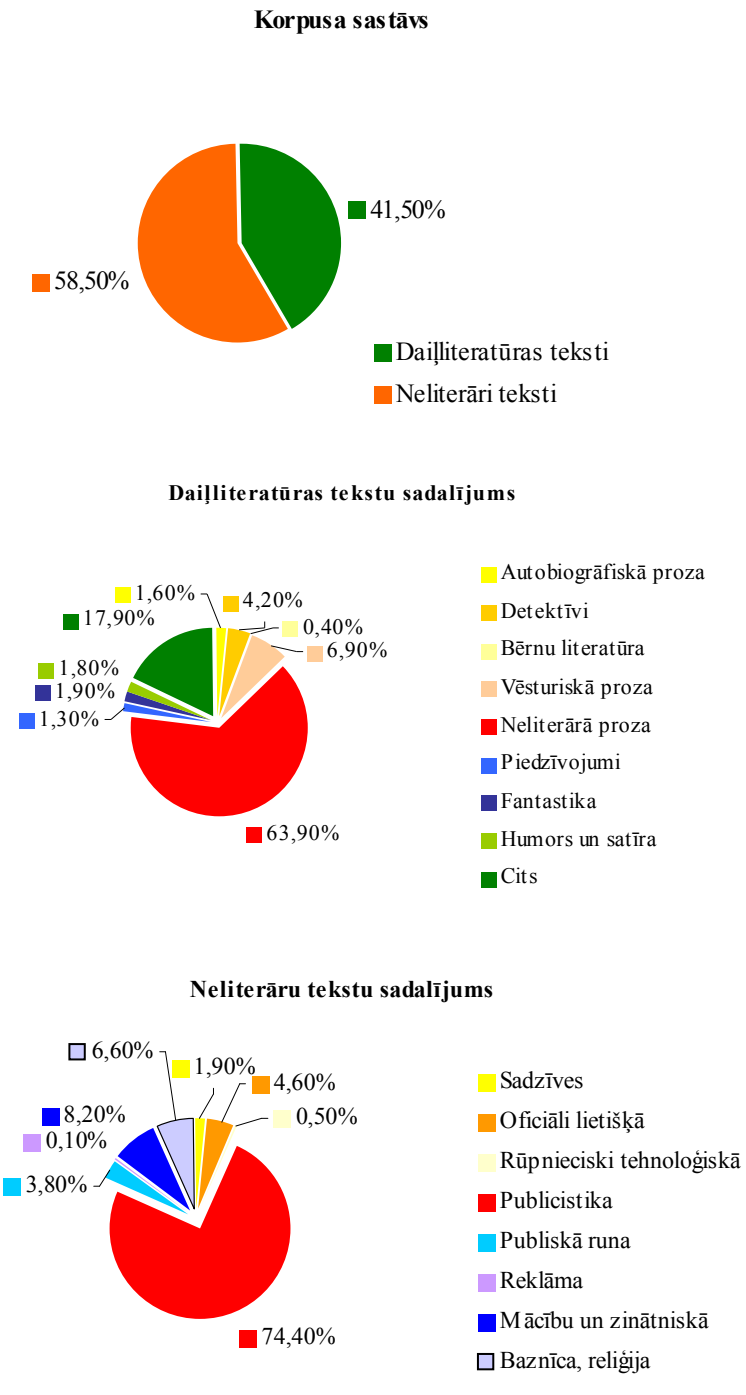
[22] http://www.natcorp.ox.ac.uk/what/garside_allc.html – skatīts 04.07.2005.; http://www.comp.lancs.ac.uk/computing/research/ucrel/claws – skatīts 04.07.2005.

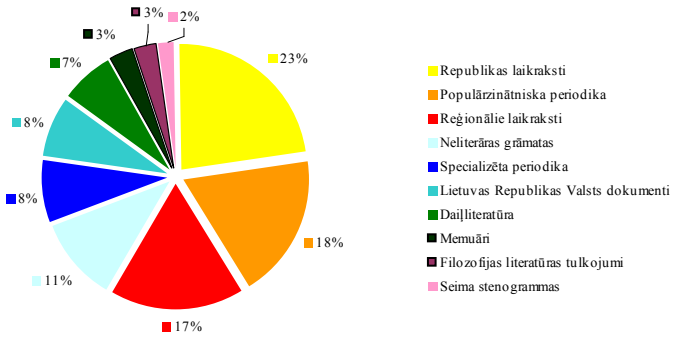
<table>
<tr><td>Autortiesības</td><td colspan="3">Sinhroniskajā korpusā iekļauti teksti, ko pēc attiecīgas vienošanās ar tekstu īpašniekiem drīkst izmantot tikai pētniecības nolūkos. Internetā publiski pieejami ir 20 tūkstoši vārdlietojumu bez morfoloģiskā marķējuma. Pilns korpuss un lietojumu funkcionalitāte pieejama bez maksas pēc attiecīga līguma noslēgšanas, kurā potenciālais korpusa lietotājs apņemas neizmantot no korpusa iegūtos datus komerciāliem nolūkiem.</td></tr>
<tr><td>Finansējums</td><td colspan="3">Česká národní banka, Komerční banka, Nakladatelství Lidové noviny, Mladá fronta DNES</td></tr>
<tr><td>Sadalījums</td><td colspan="3"> 
</td></tr>
<tr><td rowspan="5">Marķējums</td><td>Līmeņi</td><td></td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="4">Sintakse – pēc Penn Treebank modeļa[23]
SGML formāts</td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td>Morfoloģija</td><td>X</td></tr>
<tr><td>Sintakse</td><td>X</td></tr>
<tr><td rowspan="2">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Meklēšanas programmatūra Bonito[24]</td><td colspan="2">Vārdu un frāžu meklēšana, izmantojot regulārās izteiksmes; meklēšana pēc morfoloģiskās informācijas; parametrizējams konkordanču apkaimes garums; meklēšanas rezultātu kārtošana un saglabāšana; darbs ar metainformāciju</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Grieķu valodas korpuss (Hellēņu nacionālais korpuss)</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://hnc.ilsp.gr/en</td></tr>
<tr><td>Tips</td><td colspan="3">Mūsdienu tekstu korpuss</td></tr>
<tr><td>Apjoms</td><td colspan="3">34 milj. vārdlietojumu</td></tr>
</table>

[23] http://www.cis.upenn.edu/~treebank – skatīts 05.07.2005.
[24] http://nlp.fi.muni.cz/projects/bonito – skatīts 05.07.2005.

Autortiesības	Teksti iegūti no autortiesību īpašniekiem tikai pētniecības nolūkiem. Korpuss pieejams internetā. Pilnas korpusa iespējas reģistrētiem lietotājiem pieejamas par maksu tikai pētniecības nolūkiem. Bez maksas var izmantot ierobežotas meklēšanas iespējas.	
Programmatūra	**Rīks**	**Īpašības**
	Meklēšana	Meklēšana pēc vārdformas, lemmas vai vārdšķiras; var norādīt maksimālo attālumu tekstā starp vairākām meklējamām vienībām; meklēšana, izmantojot šablonus. Iespējams izgūt apakškorpusu, kas ir lietots iepriekš; var saglabāt meklēšanas rezultātus. Konkordances konteksta garuma izvēle. Meklēšanas apgabala ierobežošana, norādot metainformāciju. Meklēšana tiek nodrošināta ar morfoloģiskās vārdnīcas palīdzību. Tā satur: leksiskos datus: lemmas, celmi, galotnes, nelokāmas formas; locīšanas datus: locīšanas paradigmas; iezīmes: vārdšķiras un iezīmju kopums, kas tiek izmantots vārdformu raksturošanai morfosintaktiskā līmenī. Nešķir homoformas.
	Statistika	Tiek piedāvāta informācija par vārdformu, lemmu un vārdšķiru biežumu, kā arī dati par 100 un 1000 biežāk lietotajām vārdformām, lemmām un vārdšķirām
Nosaukums	**Horvātu nacionālais korpuss**	
Tīmekļa vietne	http://www.hnk.ffzg.hr	
Tips	Sinhronisks mūsdienu valodas korpuss. Plānoti arī citi korpusi.	
Apjoms	30 milj. vārdlietojumu mūsdienu horvātu valodas korpuss	
Izmantošana	Korpuss paredzēts, lai nodrošinātu plašākas horvātu valodas pētniecības iespējas	
Autortiesības	Korpuss brīvi pieejams internetā	
Programmatūra	**Rīks**	**Īpašības**
	Meklēšana (*Bonito*)	Sk. „Čehu nacionālais korpuss”
	Statistika	Pieejams 200 biežāk lietoto vārdu saraksts
Nosaukums	**Īru valodas nacionālais korpuss**	
Tīmekļa vietne	http://www.ite.ie/corpus	

<table>
<tr><td>Tips</td><td colspan="3">Mūsdienu valodas korpuss, kurā ievietoti rakstītie teksti īru valodā</td></tr>
<tr><td>Apjoms</td><td colspan="3">Apmēram 30 milj. vārdlietojumu. Korpuss tiek papildināts.</td></tr>
<tr><td>Izmantošana</td><td colspan="3">Korpuss izmantojams mūsdienu īru rakstītās valodas pētīšanai, vārdnīcu sastādīšanai</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Korpuss ir komerciāls produkts, ko var iegādāties CD formā par attiecīgu samaksu (250 EUR) komerciālai lietošanai un pētniecībai (50 EUR). Nākotnē plānots korpusu padarīt daļēji pieejamu arī internetā.</td></tr>
<tr><td>Finansējums</td><td colspan="3">Korpuss tapis Eiropas Savienības finansētā projekta PAROLE ietvaros (1996 – 1999). Tam piešķirts arī valsts finansējums.</td></tr>
<tr><td>Sadalījums</td><td colspan="3">Korpusā ievietoti rakstītie teksti, sākot no 1970. gada. 80% tekstu ir sarakstīti, sākot no 1980. gada.
 
</td></tr>
<tr><td rowspan="4">Marķējums</td><td>Līmeņi</td><td></td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="3">TEI vadlīnijas
SGML formāts</td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td>Morfoloģija</td><td>26%</td></tr>
<tr><td rowspan="2">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Statistika</td><td colspan="2">Pieejams 300 biežāk lietoto vārdu rādītājs.</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Krievu valodas nacionālais korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://www.ruscorpora.ru</td></tr>
<tr><td>Tips</td><td colspan="3">Krievu valodas nacionālais korpuss pamatā ir mūsdienu valodas korpuss, bet tajā plānots iekļaut arī senos tekstus. Drīzumā plānots interneta mājas lapā ievietot arī nelielu krievu-angļu paralēlo korpusu. Korpusam ir runas un tekstu daļa.
Mūsdienu literārās valodas korpusam ir divi apakškorpusi: agrīno tekstu korpuss (19. gs. sāk. – 20. gs. vidus) un mūsdienu tekstu korpuss (20. gs. vidus – 21. gs. sāk.); tīmeklī pieejams tikai mūsdienu tekstu korpuss.</td></tr>
</table>

Apjoms	Apmēram 35 milj. vārdlietojumu. Korpuss tiek papildināts, plānotais korpusa apjoms – 200 miljoni vārdlietojumu.
Izmantošana	Valodas leksikas un gramatikas (arī akcentoloģijas) zinātnisku pētījumu nodrošinājums; valodas pārmaiņu konstatēšana salīdzinoši nelielā laika posmā
Autortiesības	Pētniecības nolūkiem korpuss brīvi pieejams tīmeklī, tā tekstus nedrīkst pilnībā lasīt, kopēt
Finansējums	2003.–2005. gads – Krievijas Humanitārā zinātniskā fonda grants
Sadalījums	

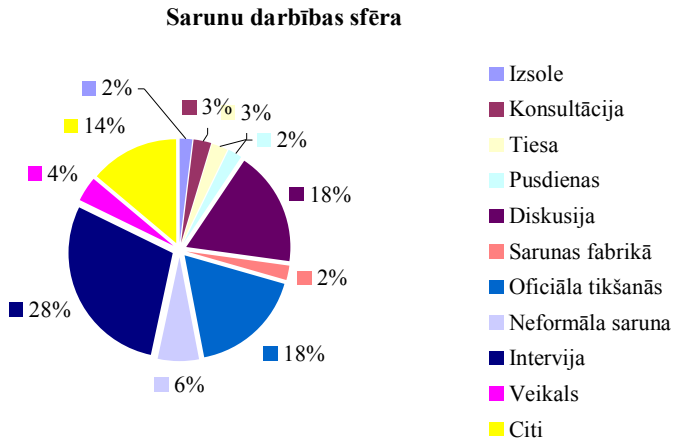
<table>
<tr><td rowspan="5">Marķējums</td><td colspan="2">Līmeņi</td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="4"></td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td>Morfoloģija</td><td>26%</td></tr>
<tr><td>Semantika</td><td>X</td></tr>
<tr><td>Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td></td><td>Meklēšana</td><td colspan="2">Korpusā var meklēt pēc vārdformas vai teksta fragmenta (ciparus, pieturzīmes u. c. simbolus nevar meklēt), kā arī pēc gramatiskām vai semantiskām pazīmēm. Parametrus var kombinēt. Meklēšanā var izmantot arī regulāras izteiksmes.</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Lietuviešu valodas tekstu korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://donelaitis.vdu.lt</td></tr>
<tr><td>Tips</td><td colspan="3">Sinhronisks mūsdienu literārās valodas korpuss</td></tr>
<tr><td>Apjoms</td><td colspan="3">102 milj. vārdlietojumu (2002. g.)</td></tr>
<tr><td>Izmantošana</td><td colspan="3">Nodrošināt plašākas lietuviešu valodas pētniecības iespējas</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Meklēšana korpusā brīvi pieejama internetā. Korpusa veidotāji piedāvā veikt papildu statistisko izpēti vai uzzināt precīzākus tekstu avotus.</td></tr>
<tr><td>Finansējums</td><td colspan="3">Lietuvas Valsts zinātnes un studiju fonds; Lietuviešu valodas valsts komisija</td></tr>
<tr><td>Sadalījums</td><td colspan="3">Korpusa sadalījums
 
</td></tr>
<tr><td rowspan="2">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Meklēšana</td><td colspan="2">Vārda vai vārda daļas meklēšana, izmantojot šablonus; meklēšana noteiktā kontekstā; meklēšana visā korpusā vai noteiktā tekstu tipā (-os)</td></tr>
</table>

	Konkordance	Iespējamais konkordances konteksta garums 30 – 300 zīmes
	Statistika	Var iegūt meklējamā vārda lietojumu skaitu un kopējo vārdu skaitu visā korpusā vai izvēlētajā tekstu tipā; ir iespēja meklēt vārdu savienojumus – tiek dots savienojums un tā relatīvais biežums. Plašāku statistisko informāciju var iegūt, nosūtot individuālu pieprasījumu korpusa veidotājiem.

Nosaukums	**Poļu valodas nacionālais korpuss**		
Tīmekļa vietne	http://korpus.ia.uni.lodz.pl		
Tips	Mūsdienu tekstu un runas korpuss (testa versija)		
Apjoms	85 milj. vārdlietojumu (tostarp 667 tūkst. runas vārdlietojumu)		
Autortiesības	Korpusa testa versija brīvi pieejama internetā		
Marķējums	**Līmeņi**		**Vadlīnijas/Formāti**
	Metadati	X	
Programmatūra	**Rīks**	**Īpašības**	
	Meklēšana	Meklēšana pēc vārda vai tā daļas, frāzes, teikuma (arī ar regulārām izteiksmēm; var norādīt, kuri vārdi nedrīkst būt meklētajā teikumā); var izvēlēties korpusa sadaļu (runa, teksts), teikuma funkcionālo veidu (stāstījuma, izsauksmes, jautājuma, nezināms), teksta avota veids (grāmata, laikraksts u. tml.), publicēšanas laiku. Var norādīt konteksta garumu, kārtot pēc vārda vai frāzes, kreisā, labā konteksta, avota. Meklēšanu var ierobežot ar metainformācijas.	
	Statistika	Tiek piedāvāts biežāk lietoto vārdu saraksts; biežāk lietotie vārdu savienojumi ar meklējamo vārdu; informācija par to, cik bieži divi meklējamie vārdi parādās kopā	

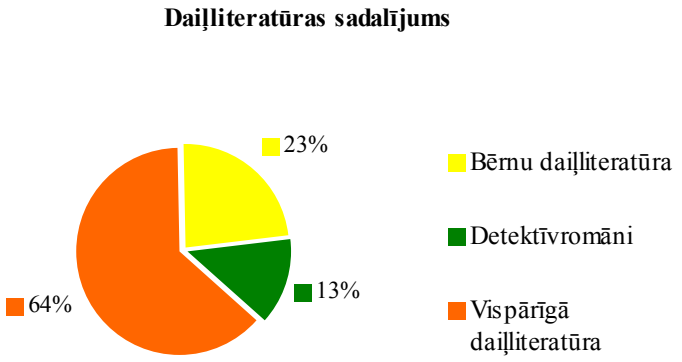
2.1.2. *Runas korpusi*

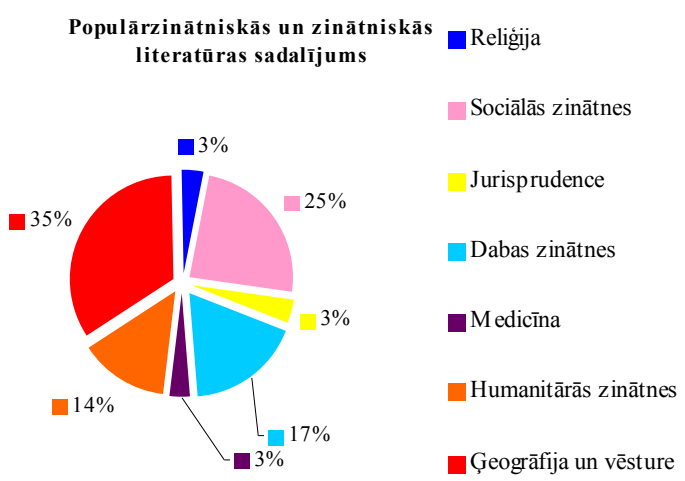
Nosaukums	**Zviedru runātās valodas korpuss**
Tīmekļa vietne	http://www.ling.gu.se/projekt/old_tal/SLcorpus.html
Tips	Runātās valodas korpuss
Apjoms	2,5 milj. vārdlietojumu, 172 h ieraksta. Korpuss tiek papildināts

<table>
<tr><td>Izmantošana</td><td colspan="3">Korpuss izmantojams zviedru runātās valodas leksikas un fonētikas pētījumiem. Īpaša uzmanība pievērsta valodas lietojumam atšķirīgās darbības sfērās.</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Korpuss nav brīvi pieejams internetā. Piekļūt tam var, iegūstot lietotājvārdu un paroli, ko piešķir tā veidotāji.</td></tr>
<tr><td>Sadalījums</td><td colspan="3"> 
</td></tr>
<tr><td rowspan="5">Marķējums</td><td colspan="2">Līmeņi</td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="4">Transkripcija pēc MSO (Modified Standard Orthography) standarta.</td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td>Morfoloģija</td><td>X</td></tr>
<tr><td>Transkripcija</td><td>X</td></tr>
<tr><td rowspan="5">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Meklēšana</td><td colspan="2">Meklēšana iespējama pēc vārda, vārdu savienojuma vai frāzes. Rezultāti tiek atspoguļoti konkordancē ar nosakāmu konteksta garumu, ar tiešu piekļuvi transkribētam tekstam.</td></tr>
<tr><td>Statistika</td><td colspan="2">Iespējams uzzināt vārdlietojumu, paužu, uzsvaru biežumu. Var tikt izrēķināta procentuālā attiecība vārdiem un pauzēm, uzsvērtajiem un neuzsvērtajiem vārdiem u. c.</td></tr>
<tr><td>TRACTOR[25]</td><td colspan="2">Transkripcijas kodēšanas rīks</td></tr>
<tr><td>TRASA[26]</td><td colspan="2">Rīks dažāda veida statistikas iegūšanai</td></tr>
<tr><td>MTTextEditor[27]</td><td colspan="2">Rīks transkripcijas pārbaudei</td></tr>
</table>

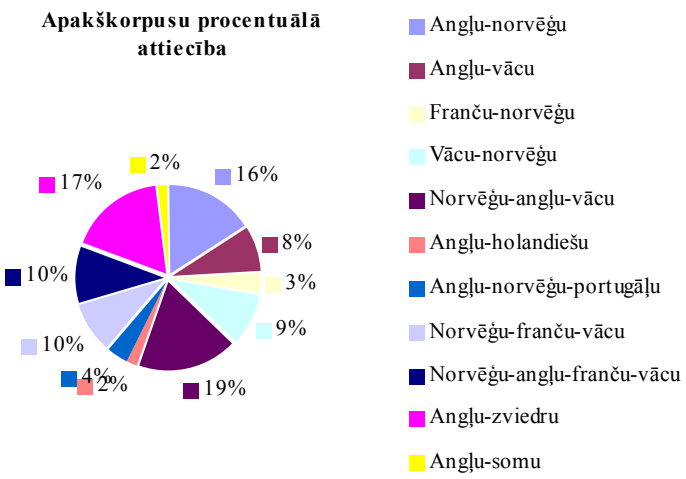
[25] http://www.ling.gu.se/~sl/tractor.html – skatīts 12.07.2005.
[26] http://www.ling.gu.se/~leifg/doc/trasa08.pdf – skatīts 13.07.2005.
[27] http://www.ling.gu.se/~mgunnar/gtseditor – skatīts 13.07.2005.

2.2. Speciālie korpusi

Nosaukums	Norvēģu-angļu paralēlais korpuss
Tīmekļa vietne	http://www.hf.uio.no/iba/prosjekt
Tips	Sinhronisks paralēlais tekstu korpuss. Tajā iekļauti paralēli norvēģu un angļu teksti, to izraudzīšanās princips – izvēlēties pēc iespējas nesenākā laikā rakstītus tekstus, lai korpuss tiktu izveidots, cik vien iespējams, viendabīgs.
Apjoms	2,6 milj. vārdlietojumu. Korpuss sastāv no 50 tekstu paraugiem, katrā no tiem 10 000 līdz 15 000 vārdlietojumu katrā valodā un to tulkojumi otrā valodā.
Izmantošana	Korpuss tiek izmantots sastatāmās valodniecības un tulkošanas pētījumiem. Korpusa pilnveidošana un tālāka attīstība turpinās. Uz šī korpusa pamata izveidots daudzvalodu tulkošanas korpuss un daudzvalodu korpuss.
Autortiesības	Korpusā ievietoti tikai tie teksti, kuru autori rakstiski piekrituši, ka atļauj izmantot tekstus valodnieciskas izpētes mērķiem. Korpusa lietošanu un uzturēšanu regulē Norvēģijas autoru un tulkotāju asociācijas noteikumi. Korpusu var lietot tikai zinātniskiem nolūkiem. Komerciālos nolūkos korpusu lietot aizliegts. Piekļuve korpusam atļauta tikai tām iestādēm, kuras saņēmušas atļauju, ko parakstījuši autortiesību turētāji. Tās ir: Oslo universitātes Britu un amerikāņu studiju departaments un Bergenas universitātes Norvēģijas humanitāro zinātņu informātikas centrs, korpusam var piekļūt arī studenti, kas saistīti ar šīm iestādēm.
Finansējums	Norvēģu-angļu paralēlais korpuss tiek veidots projekta *SPRIK (Languages in Contrast)* ietvaros. Tā finansētājs ir Norvēģijas Pētniecības padome.
Sadalījums	

<table>
<tr><td></td><td colspan="3"> 
</td></tr>
<tr><td rowspan="3">Marķējums</td><td>Līmeņi</td><td></td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="2">TEI vadlīnijas</td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td rowspan="3">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Translation Corpus Aligner[28]</td><td colspan="2">Tulkošanas korpusa sastatītājs, kas sastata divvalodu tekstu teikumu līmenī; atbalsta dažādu valodu pārus</td></tr>
<tr><td>Translation Corpus Explorer</td><td colspan="2">Korpusa tekstu meklēšanas un pārlūkošanas rīks</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Oslo daudzvalodu korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://www.hf.uio.no/iba/OMC/English/index_e.html</td></tr>
<tr><td>Tips</td><td colspan="3">Daudzvalodu korpuss</td></tr>
<tr><td>Apjoms</td><td colspan="3">16 milj. vārdlietojumu. Korpuss sastāv no 11 apakškorpusiem.</td></tr>
<tr><td>Izmantošana</td><td colspan="3">Korpuss izmantojams valodu apguvei, studijām un tulkošanai</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Oslo daudzvalodu korpuss pieejams internetā, taču piekļuve iespējama tikai pēc elektroniskas reģistrācijas veidlapas aizpildīšanas. Veidlapā tiek prasīts norādīt pieteicēja vārdu, iestādi, kurā viņš strādā/mācās, adresi un e-pasta adresi. Pieteicējam jāapstiprina, ka korpuss netiks izmantots komerciāliem mērķiem, korpusa materiāli netiks kopēti un izplatīti ārpus Oslo un Bergenas universitātes. Jānorāda arī korpusa izmantošanas mērķis.</td></tr>
<tr><td>Finansējums</td><td colspan="3">Projektu finansē Norvēģijas Pētniecības padome</td></tr>
<tr><td>Sadalījums</td><td colspan="3">Korpuss sastāv no vairākiem apakškorpusiem, tādējādi Oslo daudzvalodu korpuss būtībā ir korpusu apkopojums, kuri ir izveidoti pēc vienotas sistēmas. Korpusā iekļauti divu veidu apakškorpusi: divvalodu un tulkojamie.</td></tr>
</table>

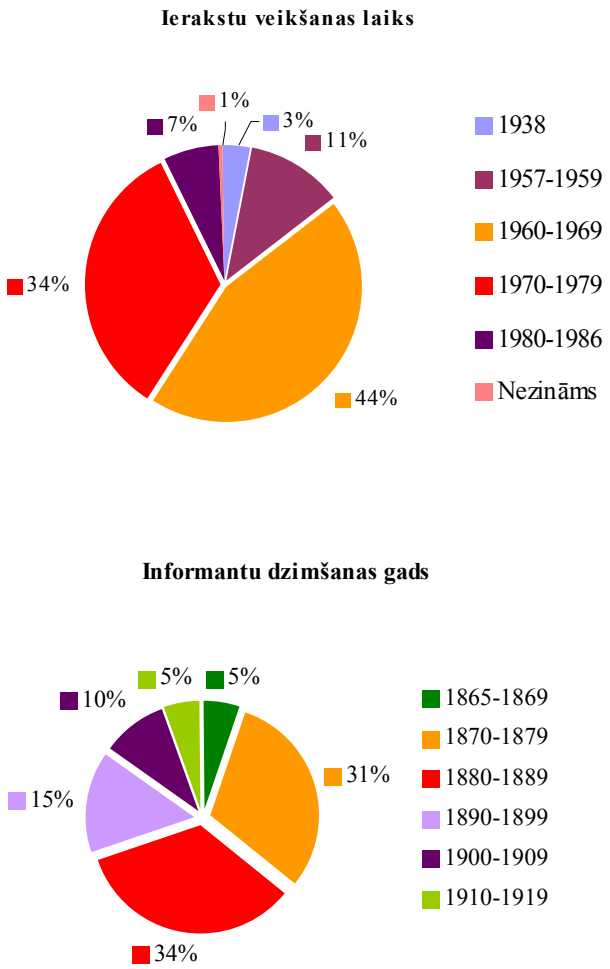
[28] http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html#_Toc445194185 – skatīts 08.07.2005.

<table>
<tr><td></td><td colspan="3"> 
</td></tr>
<tr><td rowspan="3">Marķējums</td><td>Līmeņi</td><td></td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="2">TEI vadlīnijas</td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td rowspan="3">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Translation Corpus Aligner</td><td colspan="2" rowspan="2">Sk. „Norvēģu-angļu paralēlais korpuss”.</td></tr>
<tr><td>Translation Corpus Explorer</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Bergenas Londonas pusaudžu valodas korpuss – COLT (Britu nacionālā korpusa daļa)</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://torvald.aksis.uib.no/colt</td></tr>
<tr><td>Tips</td><td colspan="3">13 – 17 gadus vecu pusaudžu runātās valodas korpuss; savākts 1993. gadā</td></tr>
<tr><td>Apjoms</td><td colspan="3">500 tūkst. vārdlietojumu, 50 stundu ieraksta</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Pieejams par maksu CD formā – tajā iekļauti skaņu faili un transkribēti teksti</td></tr>
<tr><td rowspan="3">Marķējums</td><td>Līmeņi</td><td></td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Morfoloģija</td><td>X</td><td rowspan="2"></td></tr>
<tr><td>Fonētiskā transkripcija</td><td>X</td></tr>
<tr><td rowspan="2">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Statistika</td><td colspan="2">1000 biežāk lietoto vārdu saraksts</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Starptautiskais angļu valodas mācību korpuss (ICLE)</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm</td></tr>
<tr><td>Tips</td><td colspan="3">Studentu (valodas apguvēju) mācību korpuss</td></tr>
<tr><td>Apjoms</td><td colspan="3">2 milj. vārdlietojumu</td></tr>
</table>

<table>
<tr><td>Izmantošana</td><td colspan="3">Korpuss ir empīrisks avots plaša mēroga salīdzinošajiem valodas pētījumiem, kā arī pētījumiem par klasiskajām kļūdām svešvalodu (šajā gadījumā – angļu valodas) mācīšanā, kas noder gan pārbaudes darbu izveidošanā, gan arī ļauj izveidot mācību programmu, ņemot vērā tipiskās kļūdas un pievēršot tām lielāku uzmanību</td></tr>
<tr><td>Autortiesības</td><td colspan="3">2002. gadā korpuss tika izdots CD formātā kopā ar rokasgrāmatu, kurā sniegta informācija par tā struktūru un angļu valodu tajās valstīs, kuru valodas pārstāvētas korpusā. Korpuss CD formātā ir pieejams par maksu. Korpuss pieejams arī tajās iestādēs, kas piedalās tā veidošanā. Korpuss izmantojams tikai akadēmiskiem mērķiem.</td></tr>
<tr><td>Sadalījums</td><td colspan="3">Korpuss sastāv no tekstiem – sacerējumiem – ko angliski rakstījuši studenti, kas apgūst angļu valodu, turklāt viņu valodas zināšanu līmenis ir salīdzinoši augsts. Katrs teksts ir apmēram 500 – 1000 vārdu garš. Sacerējumi ir par dažādu, taču noteiktu tematiku. Kopumā korpusā iekļauto tekstu autoriem ir 19 dažādas dzimtās valodas, kā rezultātā izveidoti ir 19 apakškorpusi: bulgāru, portugāļu (Brazīlijas), ķīniešu, čehu, holandiešu, somu, franču, vācu, itāliešu, japāņu, norvēģu, lietuviešu, poļu, portugāļu, krievu, spāņu, dienvidafrikāņu (setsvanu), zviedru, turku.</td></tr>
<tr><td rowspan="5">Marķējums</td><td>Līmeņi</td><td></td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="4">Projekta ietvaros izveidota ICE iezīmju valoda kļūdu un kļūdu labojumu marķēšanai</td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td>Morfoloģija</td><td>X</td></tr>
<tr><td>Sintakse</td><td>X</td></tr>
<tr><td rowspan="4">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>ICE Markup Assistant</td><td colspan="2">Programmatūra kļūdu un kļūdu labojumu marķēšanai</td></tr>
<tr><td>TOSCA Tagger</td><td colspan="2">Morfoloģiskais analizators/marķētājs</td></tr>
<tr><td>TOSCA Parser</td><td colspan="2">Sintaktiskais analizators/marķētājs</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Somijas zviedru valodas tekstu korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://www.nord.helsinki.fi/fisc/presseng.html</td></tr>
<tr><td>Tips</td><td colspan="3">Mūsdienu zviedru valodas rakstītās valodas korpuss, teksti iegūti no zviedru valodā publicētajiem tekstiem Somijā deviņdesmitajos gados. Iekļauta arī neliela daļa runātās valodas.</td></tr>
<tr><td>Apjoms</td><td colspan="3">2,5 milj. vārdlietojumu</td></tr>
<tr><td>Izmantošana</td><td colspan="3">Korpuss izmantojams Somijā runātās zviedru valodas pētīšanā</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Korpuss pieejams jebkuram. Lietotājam jāsazinās ar korpusa veidotājiem, viņam tiek piešķirts lietotāja vārds un parole.</td></tr>
</table>

<table>
<tr><td>Sadalījums</td><td colspan="3">Korpusā iekļauti teksti no šādiem avotiem: prese; daiļliteratūra; neliterārā proza; oficiālie dokumenti (likumu teksti, administratīvie teksti); ikdienas sarunas. Šie teksti publicēti deviņdesmitajos gados.</td></tr>
<tr><td rowspan="3">Marķējums</td><td colspan="2">Līmeņi</td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Morfoloģija</td><td>X</td><td rowspan="2">TEI vadlīnijas</td></tr>
<tr><td>Fonētiskā transkripcija</td><td>X</td></tr>
<tr><td rowspan="2">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>SWETWOL</td><td colspan="2">Zviedru valodas morfoloģiskais analizators</td></tr>
<tr><td>Nosaukums</td><td colspan="3">22 valodu telefona sarunu korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://cslu.cse.ogi.edu/corpora/22lang</td></tr>
<tr><td>Tips</td><td colspan="3">Telefona sarunu korpuss 22 valodās. Iekļautās valodas: austrumarābu valoda, kantoniešu, čehu, farsi, franču, vācu, hindi, ungāru, japāņu, korejiešu, malajiešu, mandarīnu, itāliešu, poļu, portugāļu, krievu, spāņu, zviedru, suahili, tamiliešu, vjetnamiešu, angļu valoda.</td></tr>
<tr><td>Apjoms</td><td colspan="3">Kopējais skaņu failu garums – 99 stundas
Ortogrāfiski transkribēto skaņu failu garums – 40 stundas</td></tr>
<tr><td>Izmantošana</td><td colspan="3">Korpuss paredzēts gan akadēmiskiem, gan komerciāliem runas pētījumiem, gan arī kā palīglīdzeklis runas sistēmu izveidē</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Korpuss par maksu pieejams gan pētnieciskiem, gan komerciāliem mērķiem (jānoslēdz atšķirīgi līgumi)</td></tr>
<tr><td>Sadalījums</td><td colspan="3">Korpusā ietilpst noteikta fiksēta leksika, kā arī brīvas sarunas. Leksika iegūta, runātājiem atbildot uz noteiktiem jautājumiem[29]</td></tr>
<tr><td rowspan="3">Marķējums</td><td colspan="2">Līmeņi</td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="2"></td></tr>
<tr><td>Fonētiskā transkripcija</td><td>X</td></tr>
<tr><td rowspan="2">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>Speech View</td><td colspan="2">Ļauj apskatīt spektrogrammas, labot skaņu failus u. tml.</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Igauņu valodas dialektu korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://www.murre.ut.ee</td></tr>
<tr><td>Tips</td><td colspan="3">Dialektu korpuss, kurš sastāv no runātā teksta (skaņu ierakstiem) un to atšifrējumiem</td></tr>
<tr><td>Apjoms</td><td colspan="3">600 tūkst. vārdlietojumu. Darbs pie korpusa papildināšanas un pilnveidošanas turpinās.</td></tr>
</table>

[29] http://cslu.cse.ogi.edu/corpora/22lang/protocol.html – skatīts 18.07.2005.

<table>
<tr><td>Izmantošana</td><td>Korpuss izmantojams dialektoloģijas studijām, dialektu pētniecībai, dialektu salīdzināšanai un dialektu vārdnīcu veidošanai</td></tr>
<tr><td>Autortiesības</td><td>Korpuss ir brīvi pieejams internetā. Nekādi īpaši nosacījumi tā izmantošanai nav minēti.</td></tr>
<tr><td>Sadalījums</td><td> 
</td></tr>
<tr><td>Marķējums</td><td><table>
<tr><th>Līmeņi</th><th></th><th>Vadlīnijas/Formāti</th></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td><td rowspan="2">TEI vadlīnijas
XML formāts</td></tr>
<tr><td>Morfoloģija</td><td>25%</td></tr>
</table></td></tr>
<tr><td>Programmatūra</td><td><table>
<tr><th>Rīks</th><th>Īpašības</th></tr>
<tr><td>Mark</td><td>Daļēji automatizēts tekstu strukturālās un morfoloģiskās marķēšanas rīks</td></tr>
<tr><td>Meklēšana</td><td>Iespējams norādīt dažādus meklēšanas kritērijus</td></tr>
</table></td></tr>
<tr><td>Nosaukums</td><td>Dikensa korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td>http://www.ims.uni-stuttgart.de/projekte/CQPDemos/CQPDemo/frames-cqp.html</td></tr>
<tr><td>Tips</td><td>Viena autora valodas korpuss</td></tr>
</table>

<table>
<tr><td>Apjoms</td><td colspan="3">3,4 milj. vārdlietojumu</td></tr>
<tr><td>Izmantošana</td><td colspan="3">Korpuss izmantojams Čārlza Dikensa valodas pētniecībai</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Korpuss pieejams bez maksas internetā. Autortiesības attiecas uz programmrīkiem.</td></tr>
<tr><td rowspan="5">Marķējums</td><td colspan="2">Līmeņi</td><td>Vadlīnijas/Formāti</td></tr>
<tr><td>Metadati</td><td>X</td><td rowspan="4"></td></tr>
<tr><td>Iekšējā, loģiskā struktūra</td><td>X</td></tr>
<tr><td>Morfoloģija</td><td>25%</td></tr>
<tr><td>Sintakse</td><td>X</td></tr>
<tr><td rowspan="4">Programmatūra</td><td>Rīks</td><td colspan="2">Īpašības</td></tr>
<tr><td>TreeTagger[30]</td><td colspan="2">Valodneatkarīga, automatizēta vārdšķiru marķēšanas un lemmatizēšanas sistēma.</td></tr>
<tr><td>Meklēšana</td><td colspan="2">Korpusā var meklēt pēc vārda, vārda daļas, frāzes, teikuma, rindkopas. Meklējot var izvēlēties, kādā veidā lietotājs vēlas, lai tiktu atspoguļots rezultāts. Var izvēlēties apskatīt marķētu tekstu.</td></tr>
<tr><td>Konkordance</td><td colspan="2">Lietotājs var norādīt konteksta garumu konkordancē: 5 vārdi, teikums, 2 teikumi vai rindkopa. Konkordancē atlasītos fragmentus ir iespējams sakārtot vairākos veidos, piemēram, pēc biežuma. Katram fragmentam norādīts romāns, no kura tas ņemts, un romāna nodaļa.</td></tr>
<tr><td>Statistika</td><td colspan="2">Korpusā iespējams iegūt meklējamā vārda, vārdu savienojuma, vārdkopas lietojumu skaitu un tā izplatību korpusā ievietotajos romānos.</td></tr>
<tr><td>Nosaukums</td><td colspan="3">Sintaktiski anotēts vācu laikrakstu korpuss</td></tr>
<tr><td>Tīmekļa vietne</td><td colspan="3">http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus</td></tr>
<tr><td>Tips</td><td colspan="3">Mūsdienu laikrakstu tekstu korpuss</td></tr>
<tr><td>Apjoms</td><td colspan="3">355 tūkst. vārdlietojumu</td></tr>
<tr><td>Autortiesības</td><td colspan="3">Pētniecības nolūkiem korpuss pieejams bez maksas pēc attiecīga līguma noslēgšanas. Korpusa izmantošanas komerciālā licence par maksu (4000 EUR).</td></tr>
<tr><td>Finansējums</td><td colspan="3">Sārlandes universitāte; Deutsche Forschungsgemeinschaft</td></tr>
</table>

[30] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html – skatīts 1.07.2005.

Marķējums	Līmeņi		Vadlīnijas/Formāti
	Morfoloģija	25%	
	Sintakse	X	
Programmatūra	**Rīks**	**Īpašības**	
	Meklēšana		
	Konkordance		

2.3. Kopsavilkums

2.3.1. *Reprezentativitāte*

Par dažādiem korpusiem pieejams atšķirīgs informācijas daudzums. Piemēram, par Britu nacionālā korpusa un Krievu valodas nacionālā korpusa sastāvu ir apkopota plaša statistiskā informācija (par tekstu, autoru, mērķauditoriju u. tml.), korpuss statistiski raksturots no dažādiem viedokļiem. Savukārt par dažu korpusu sastāvu pieejama ļoti trūcīga informācija, piemēram, Grieķu valodas korpusa tīmekļa vietnē minēts tikai viens kritērijs korpusā ievietojamo tekstu atlasē – lasītākie teksti.

Visbiežāk lielāko korpusa daļu veido laikraksti un citi periodiskie preses izdevumi (arī reģionālie). Apskatītajos korpusos (tajos, par kuriem ir pieejama attiecīga informācija) preses īpatsvars svārstās no 31,08 % (Britu nacionālais korpuss) līdz 66,5 % (Čehu nacionālais korpuss). Salīdzinoši nelielu korpusa daļu parasti aizņem daiļliteratūra – 7 % (Lietuviešu valodas tekstu korpuss) – 21,91 % (Britu nacionālais korpuss). Izņēmums šajā ziņā ir Krievu valodas nacionālais korpuss, kurā daiļliteratūras teksti ir 41,5 % – gandrīz tikpat daudz kā publicistikas tekstu šajā korpusā (43,52 %). Zinātniskā vai speciālā literatūrā korpusos veido no 4,8 % (Krievu valodas nacionālais korpuss) līdz 33,5 % (Čehu nacionālais korpuss). Pārējās korpusa daļas visbiežāk ir samērā nelielas (parasti nepārsniedz 10 %) un dažādos korpusos atšķirīgas. Starp tām mēdz būt normatīvie akti un citi dokumenti, runu teksti, reklāmas, reliģiskā literatūra u. c.

2.3.2. *Apjoms*

Visbiežāk sastopamais vispārīgo korpusu apjoms ir aptuveni 100 miljoni vārdlietojumu. Daļai korpusu apjoms ir mazāks, bet tie vēl ir tapšanas stadijā, vairumam no tiem plānots sasniegt 100 miljonu vārdlietojumu apjomu, Krievu valodas nacionālā korpusa apjoms plānots 200 miljonu vārdlietojumu.

Speciālo korpusu apjoms ir mazāks, jo šajos korpusos ievieto ierobežotu tekstu apjomu, kur svarīga ir nevis tekstu kvantitāte, bet – to specifika, piemēram, dialektoloģijas materiāls, noteiktas vecuma grupas runātie teksti vai viena autora rakstītie teksti. Visbiežāk sastopamais vārdlietojumu skaits ir ap 2 miljoniem.

Piemēram, Angļu valodas mācību korpusā – 2 miljoni vārdlietojumu, Zviedru runātās valodas korpusā ir 2,5 miljoni vārdlietojumu, līdzīgi arī Norvēģu-angļu paralēlajā korpusā – 2,6 miljoni vārdlietojumu. Savukārt Bergenas Londonas pusaudžu valodas korpusā ir tikai 500 tūkstoši vārdlietojumu un Igauņu dialektu korpusā – apmēram 600 tūkstoši vārdlietojumu. Taču gandrīz visu speciālo korpusu pilnveidošana turpinās.

2.3.3. *Marķējums*

Lielākā daļā korpusu tekstu ir marķēti, izmantojot starptautiskās TEI (*Text Encoding Iniciative*) vadlīnijas, piemēram, Britu nacionālajā korpusā, Īru valodas nacionālajā korpusā, Norvēģu-angļu paralēlajā korpusā, Somijas zviedru valodas korpusā u. c. Izplatītākais formāts marķējuma realizācijai ir SGML (*Standard Generalized Markup Language*; ISO:8879); pēdējo gadu izstrādēs SGML tiek aizstāts ar XML standartu (*eXtensible Markup Language*). Praktiski visos korpusos teksti tiek klasificēti pēc dažādām vairāk vai mazāk detalizētām pazīmēm (metainformācija). Piemēram, Čehu nacionālajā korpusā izmantotas 12 veidu pazīmes: korpusa tips, teksta tips, teksta žanrs, teksta apakšžanrs, teksta avota veids, autora dzimums, valoda, oriģinālvaloda, izdošanas gads, autora/tulkotāja vārds, lielāka teksta identifikācija.

Lielākajā daļā korpusu anotētas tiek dažādas teksta iekšējās struktūras kategorijas: virsraksts, rindkopa, teikums, nodaļa, komentāri, tiešā runa u. tml. Šim nolūkam tiek izmantoti automatizēti analīzes un marķēšanas rīki.

Gramatisko kategoriju automatizētai marķēšanai tiek izmantotas īpaši šim nolūkam izstrādātas programmas. Katrai valodai, ņemot vērā tās īpatnības, šāda programma ir jāizstrādā atsevišķi, taču eksistē arī vairākas sistēmas, kuras tiek aprakstītas, kā valodneatkarīgas vai pielāgojamas, piemēram, *TreeTagger*. Programmas marķēšanu veic automatizēti, un, tās izmantojot, pilnībā no kļūdām izvairīties nav iespējams, tāpēc vienmēr tiek norādīts, ka marķēšana notiek daļēji automatizēti.

Gandrīz visiem aplūkotajiem korpusiem, izņemot Lietuviešu valodas nacionālo korpusu un runātās valodas korpusus, ir marķētas vārdšķiras, kā arī citas gramatiskās kategorijas: dzimte, skaitlis, locījums u. tml.

Ņemot vērā korpusa veidu, var tikt marķētas arī specializēto korpusu tekstiem īpaši raksturīgas iezīmes, piemēram, Starptautiskajā angļu valodas mācību korpusā īpaši marķētas ir gramatiskās kļūdas.

Sintaktiskā analīze tiek norādīta vairākos korpusos: Čehu nacionālajā korpusā, Angļu valodas mācību korpusā, Dikensa korpusā. Arī sintaktiskās analīzes nolūkā tiek izstrādāti programmrīki darba automatizēšanai, piemēram, *TOSCA Parser* (Starptautiskais angļu valodas mācību korpuss).

2.3.4. *Lietojumrīki*

Korpusos galvenais meklēšanas programmrīks ir konkordance. Aplūkotajos korpusos bieži tiek piedāvāta iespēja izvēlēties konkordances konteksta garumu. Tas variējas no dažiem vārdiem līdz rindkopai vai pat neierobežotam konkordances konteksta garumam, kā tas ir, piemēram, Dikensa korpusā.

Korpusa lietotājiem tiek piedāvātas ļoti plašas meklēšanas iespējas. Iespējams meklēt veselu vārdu vai tā daļu, vārdu savienojumu, frāzi, teikumu vai pat rindkopu. Marķētos korpusos meklēt var arī noteiktas gramatiskās kategorijas – vārdšķiru, locījumu u. tml. Atkarībā no lietotāja vēlmēm ir iespējams meklēt noteiktā korpusa daļā, piemēram, norādot konkrētā periodā izdotus tekstus, var meklēt noteikta veida tekstos, piemēram, periodikā vai daiļliteratūrā. Atsevišķos korpusos meklēšanas rezultātus var saglabāt. Korpusā atlasīto rezultātu šķirošana arī tiek piedāvāta pēc dažādiem kritērijiem: pieminējuma biežuma, kādas noteiktas vārda gramatiskās kategorijas u. tml.

Korpusos, kuros ir dažādas kategorijas lietotāji, piemēram, lietotāji, kam ir daļēja piekļuve korpusam, un lietotāji, kam ir pilnīga piekļuve korpusam, arī meklēšanas iespējas tiek variētas atkarībā no lietotāja kategorijas. Piemēram, Grieķu valodas korpusā reģistrētie lietotāji var saglabāt meklēšanas rezultātus, savukārt nereģistrētie – nevar.

Lielākajā daļā korpusu ir iespējams iegūt dažādus statistiskos rādītājus, piemēram, attiecīgā vārda lietojumu skaitu korpusā vai noteiktā korpusa daļā. Ir iespēja meklēt stabilus vārdu savienojumus un uzzināt to biežumu korpusā vai tā daļās. Nereti pieejams arī biežāk lietoto vārdu saraksts korpusā. Šie ir vienkāršākie statistiskie rādītāji, kas parasti pieejami korpusos.

2.3.5. *Izmantošana*

Korpusu izmantošana, protams, lielā mērā atkarīga no lietotāja iecerēm un nolūka. Informācijā par korpusiem parasti tiek norādīts, ka korpuss pirmām kārtām izmantojams valodas pētniecībā. Atkarībā no korpusa veida, tas izmantojams leksikoloģijas, morfoloģijas, sintakses vai fonētikas pētījumiem. Daudzi korpusi paredzēti arī dažāda veida vārdnīcu izstrādei. Specializēto korpusu izmantošana ir šaurāka un vairāk specializēta, piemēram, Starptautiskais angļu valodas mācību korpuss paredzēts valodas apgūšanai un valodas mācīšanas programmu izstrādei, ņemot par pamatu attiecīgās valodas runātāju tipiskās kļūdas svešvalodā, kā arī pētījumiem par klasiskajām kļūdām angļu valodas apguvē. Savukārt Igauņu dialektu korpuss paredzēts igauņu valodas izlokšņu pētniecībai un dialektoloģijas vārdnīcu sastādīšanai.

2.3.6. *Autortiesības, piekļuve*

Visu korpusu (par kuriem ir attiecīgā informācija) veidotāji ir vienojušies ar tekstu devējiem – autortiesību īpašniekiem par tekstu (visbiežāk tekstu fragmentu) izmantošanu pētniecības nolūkiem.

Piekļuve dažādiem korpusiem atšķiras. Daļa korpusu (piem., Lietuviešu valodas tekstu korpuss, Poļu valodas nacionālais korpuss u. c.) ir brīvi pieejami internetā. Dažos gadījumos (Čehu nacionālais korpuss, Grieķu valodas korpuss) internetā brīvi pieejams ir samazināta apjoma vai iespēju korpuss; lai piekļūtu korpusa pilnai versijai, lietotājam jānoslēdz ar korpusa veidotājiem līgums. Reizēm (piemēram, Britu nacionālais korpuss, Īru valodas nacionālais korpuss) jebkāda piekļuve korpusam iespējama tikai pēc līguma noslēgšanas.

Korpusa izmantošanas līgumi mēdz būt dažādi. Lielākā daļa korpusu paredzēti izmantošanai tikai pētniecības mērķiem. Daļā gadījumu (piem., Čehu nacionālais korpuss, Krievu valodas nacionālais korpuss) šāda korpusa izmantošana ir bez maksas. Atsevišķi jāmin, ka šajā ziņā atšķiras Krievu valodas Nacionālais korpuss – lai to izmantotu, nav jāparaksta līgums, korpusa veidotāji aprobežojas ar to, ka korpusa mājas lapā ir ielikts paziņojums par autortiesību aizsardzību un korpusa izmantojumu. Nereti (piem., Īru valodas nacionālais korpuss, Britu nacionālais korpuss, Grieķu valodas korpuss) piekļuve korpusam arī pētniecības mērķiem ir par maksu.

Komerciāliem nolūkiem vispārīgos korpusus var izmantot tikai par maksu, šāda iespēja paredzēta ne visiem korpusiem.

Lielākā daļa korpusu pieejama tiešsaistē internetā, taču atsevišķos gadījumos (piem., Britu nacionālais korpuss, Starptautiskais angļu valodas mācību korpuss) korpuss tiek izplatīts arī CD formātā.

2.3.7. *Finansējums*

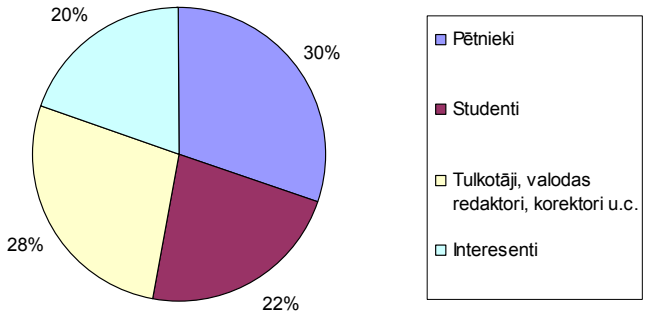
Liela daļa korpusu veidoti ar valsts iestāžu finansiālu atbalstu. Piemēram, Britu nacionālo korpusu atbalstījusi Lielbritānijas Zinātnes un tehnikas padome, Britu bibliotēka, Britu akadēmija u. c. Valsts finansiālu atbalstu saņēmis arī Čehu nacionālais korpuss, Lietuviešu valodas tekstu korpuss, Oslo daudzvalodu korpuss, Norvēģu-angļu paralēlais korpuss un Sintaktiski anotētais vācu laikrakstu korpuss.

Atsevišķi korpusi veidoti plašāku projektu ietvaros un saņēmuši Eiropas Savienības finansējumu, piemēram, Īru valodas nacionālais korpuss, kurš veidots kā Eiropas Savienības projekta PAROLE daļa no 1996. – 1999. gadam. Taču līdztekus Eiropas Savienības finansējumam šim korpusam piešķirts arī valsts finansējums.

3. Sabiedriskā aptauja

> Valodnieku un citu iespējamo korpusa lietotāju (sabiedrības) interešu un vajadzību noskaidrošanai tika organizēta aptauja, izsūtot vēstules atsevišķām organizācijām un ievietojot aptaujas anketu internetā.

Valodas korpusa koncepcijas izstrādes laikā atsevišķām organizācijām un zinātniskām iestādēm tika izsūtītas apmēram 200 anketas, kā arī ievietota anketa LU MII tīmekļa vietnē[31]. Aptaujas laikā gan elektroniski, gan arī papīra formātā tika saņemtas 76 respondentu atbildes, kuru nosacīts sadalījums ir šāds:



3.1. diagramma – respondentu sadalījums.

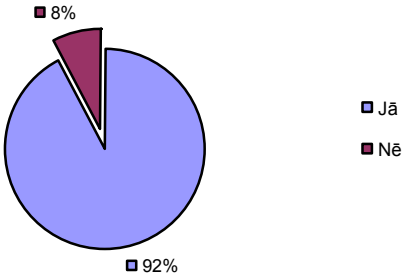
Atsaukušies ir pētnieki no dažādām Latvijas un ārvalstu zinātniskajām iestādēm, piemēram, LU Filoloģijas fakultātes, LU Pedagoģijas un psiholoģijas fakultātes, Liepājas Pedagoģijas akadēmijas, Sanktpēterburgas Valsts universitātes, Rīgas Pedagoģijas un izglītības vadības augstskolas, Hāgenas Tālmācības universitātes Rīgas studiju centra, Rēzeknes Augstskolas, E. M. Arndta universitātes Baltistikas institūta, Latviešu valodas institūta un Maksa Planka Evolucionārās antropoloģijas institūta. Studenti, kas piedalījušies aptaujā, galvenokārt ir baltu specialitātes studenti no Latvijas Universitātes. Aptaujai atsaukušies cilvēki, kuru profesionālā darbība ir saistīta ar latviešu valodu: tulkotāji (gan no lielākajām Latvijas tulkošanas aģentūrām, gan no Eiropas Parlamenta), valodas redaktori, korektori, terminologi (to starpā arī no Eiropas centrālās bankas) un Latvijas žurnālisti. Citu interesentu vidū, kuru profesionālā darbība nav saistīta ar latviešu valodu, jāmin ekonomisti, programmētāji un lektori.

Anketā tika uzdoti 12 jautājumi valodas korpusa lietotāju interešu noskaidrošanai. Turpmāk tiks apskatīti visi jautājumi un sniegta respondentu atbilžu analīze.

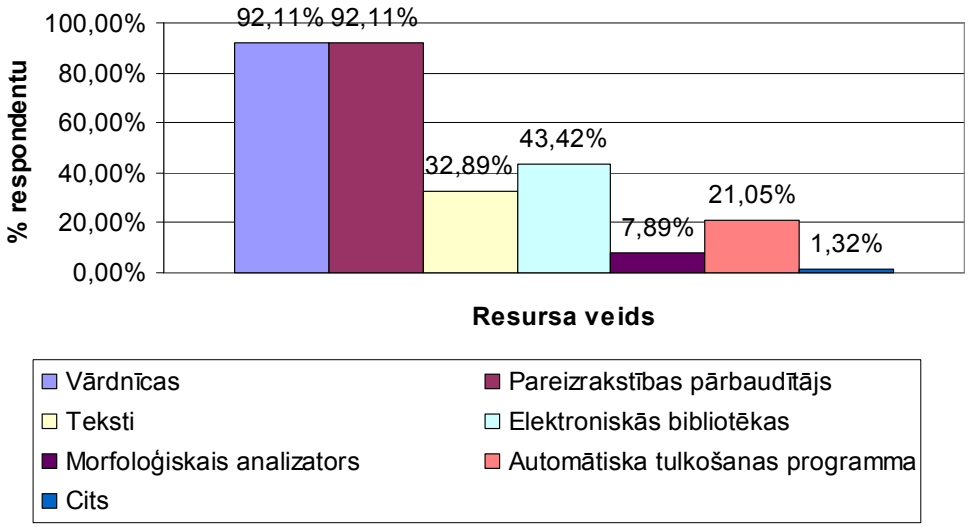
[31] http://www.ailab.lv/korpuss – skatīts 22.07.2005.

3.1. Vai Jūs izmantojat latviešu valodas elektroniskos resursus un programmrīkus?

Jā. / Nē. Ja jā, tad, lūdzu, precizējiet: pareizrakstības pārbaudītāju, vārdnīcas, tekstus, elektroniskās bibliotēkas, morfoloģisko analizatoru, automātiskās tulkošanas programmu, citu.



3.2. diagramma – latviešu valodas elektronisko resursu un programmrīku izmantojums.

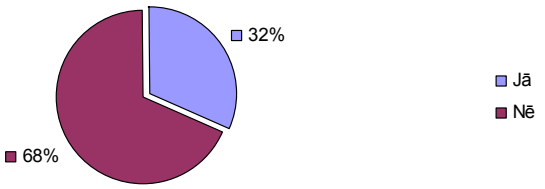


3.3. diagramma – latviešu valodas elektronisko resursu un programmrīku veidu izmantojums.

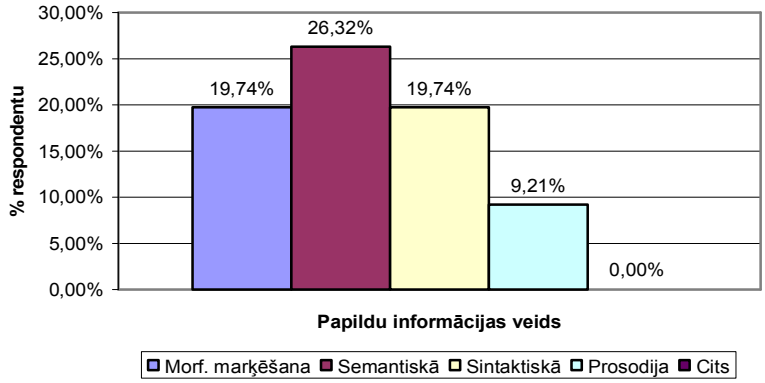
Jāpiebilst, ka 1 respondents atzina, ka ikdienas darbā izmanto programmu vārdu sadalīšanai zilbēs, tādējādi veidojot 1,32%.

3.2. Vai Jūs izmantotu elektroniskos tekstus ar papildu informāciju (marķēšanu jeb anotēšanu)?

Jā. / Nē. Ja jā, tad, lūdzu, precizējiet: morfoloģisko marķējumu (ziņas par vārdšķirām, locījumiem u. c.), semantisko marķējumu, sintaktisko marķējumu, prosodijas marķējumu (runas tekstiem).



3.4. diagramma – elektronisko tekstu ar papildu informāciju potenciālie izmantotāji.

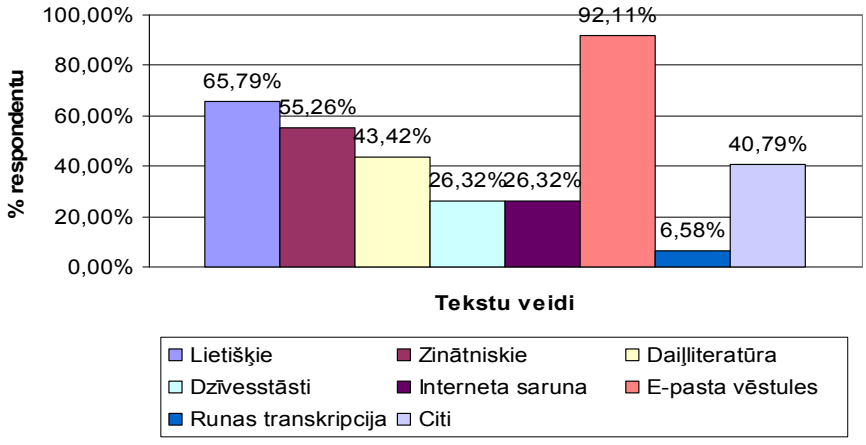


3.5. diagramma – elektronisko tekstu ar papildu informāciju veidu potenciālais izmantojums.

Vislielākā interese ir pausta par tekstu semantisko marķējumu, savukārt valodas korpusos visbiežāk sastopamais – morfoloģiskais – marķējums aptaujātos respondentus interesējis mazāk.

3.3. Ar kādiem tekstiem Jūs strādājat?

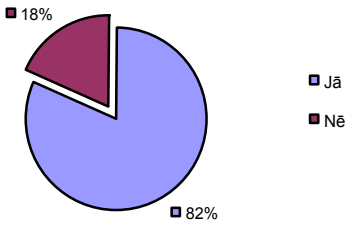
Ar lietišķajiem, zinātniskajiem (lūdzu, precizējiet, kāda joma), daiļliteratūru, dzīvesstāstiem, interneta sarunām, e-pasta vēstulēm, runas transkripciju, citiem.



3.6. diagramma – lietotāju izmantojamie tekstu veidi.

Atbildot uz šo jautājumu, visvairāk – 92,11% – respondentu atbildēja, ka strādā ar e-pasta vēstulēm. Svarīgs ir lietišķo tekstu īpatsvars, jo 65,79% respondentu strādā ar šiem tekstiem. Savukārt 40,79% izvēlējās citus tekstus: visbiežāk tika minēta publicistika, tad folkloras teksti, dialektu materiāli un senie teksti.

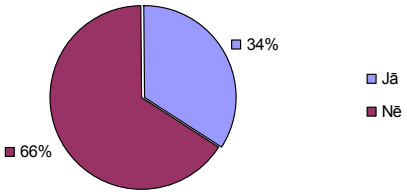
3.4. Vai Jūs strādājat ar citu valodu datiem?



3.7. diagramma – citu valodu datu izmantojums.

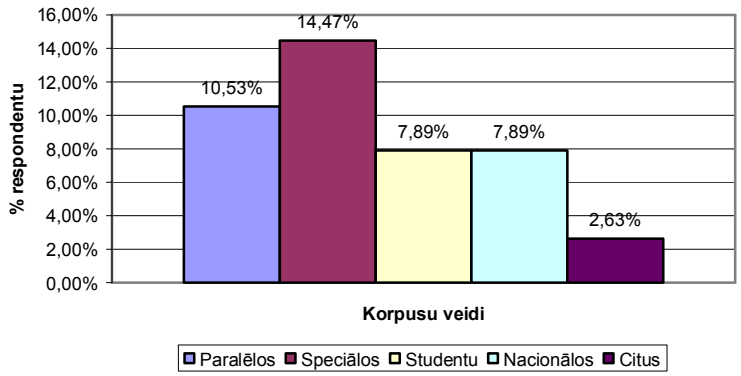
3.5. Vai Jūs izmantojat tekstu korpusus citās valodās?

Jā. / Nē. Ja jā, tad, lūdzu, precizējiet: paralēlos, speciālos, studentu, nacionālos, citus.



3.8. diagramma – tekstu korpusu citās valodās izmantojums.

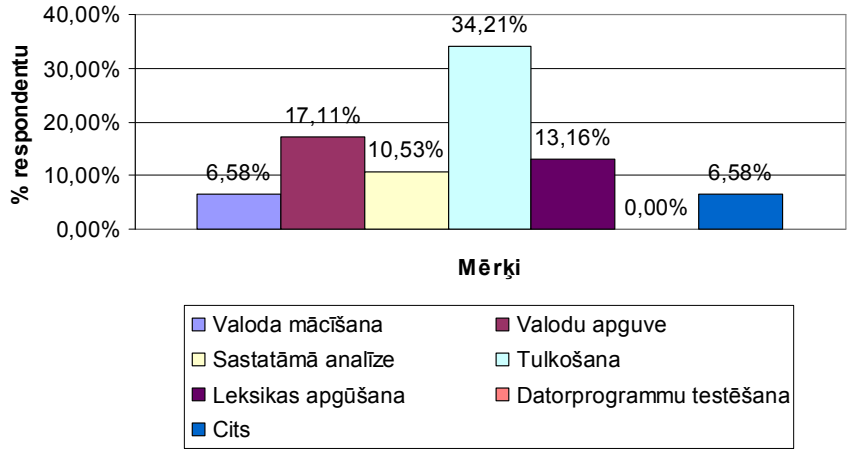
Atbildot uz 5. jautājumu, trešdaļa respondentu atbildēja, ka izmanto tekstu korpusus citās valodās, visbiežāk speciālos un paralēlos korpusus, pie citiem minēts vispārīgs korpuss.



3.9. diagramma – tekstu korpusu veidu citās valodās izmantojums.

3.6. Ja Jūs izmantojat tekstu korpusus citās valodās, nosauciet, lūdzu, mērķi:

valodu mācīšana, valodu apguve, sastatāmā analīze, tulkošana, leksikas apgūšana, datorprogrammu testēšana, cits.

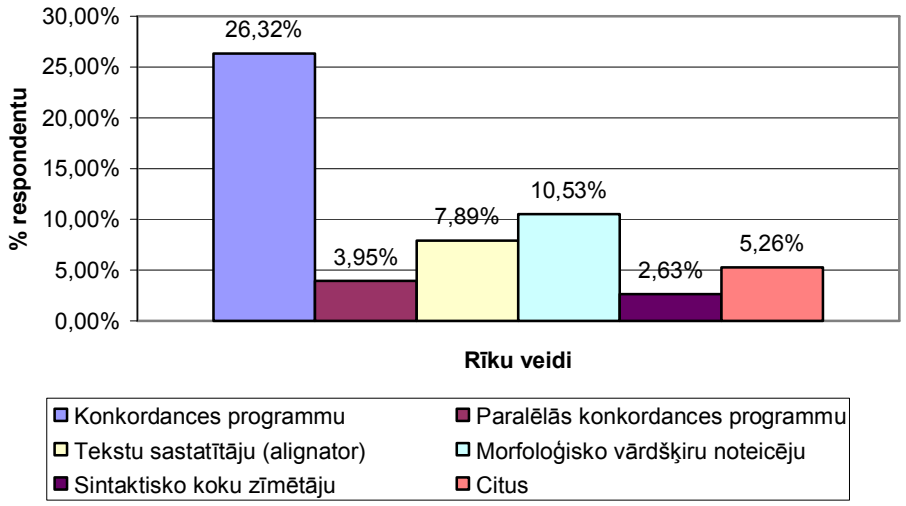


3.10. diagramma – tekstu korpusu citās valodās izmantojuma mērķis.

Visbiežāk citu valodu korpusu izmanto praktiskās darbībās – tulkošanā. Nākamais korpusu izmantošanas veids saistīts ar valodas mācīšanos. 6,58% respondentu izvēlējās citu mērķi: pētniecību (tai skaitā arī diahronijas un tipoloģijas pētījumus) un vārdnīcas izstrādi.

3.7. Vai esat lietojis kādu no šiem programmrīkiem?

Konkordances programmu (kas sameklē norādītā vārda visus lietojumus tekstā un katru no tiem novieto iekrāsotu savā rindā ekrāna vidū, uzrādot arī tekstu abās pusēs no tā), paralēlās konkordances programmu, tekstu sastatītāju (*alignator*), morfoloģisko vārdšķiru noteicēju, sintaktisko koku zīmētāju, citus.

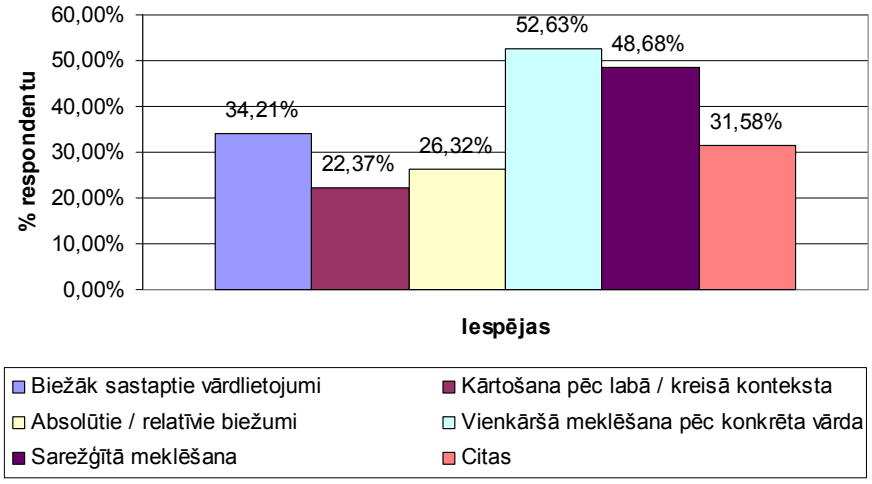


3.11. diagramma – programmrīku izmantojums.

Jāatzīst, ka lietotāju pieredze darbā ar valodas apstrādes un analīzes programmrīkiem nav ļoti liela, vispazīstamākais rīks ir konkordances programma. Tā kā daudzi respondenti citu valodu korpusus izmanto tulkošanā (3.10. diagrammā redzams, ka to par galveno mērķi atzinuši 34,21% respondentu), tad tiek lietots arī tekstu sastatītājs un paralēlās konkordances programma.

3.8. Kādas iespējas vēl Jums būtu noderīgas?

Biežāk sastaptie vārdlietojumi, kārtošana pēc labā/kreisā konteksta, absolūtie/relatīvie biežumi, vienkāršā meklēšana pēc konkrēta vārda, sarežģītā meklēšana (piem., ievadot formu *audz*, tiek piedāvāti arī piemēri ar *augsim, augām* utt.), citas.



3.12. diagramma – iespējamo programmrīku un to sniegto iespēju izmantojums.

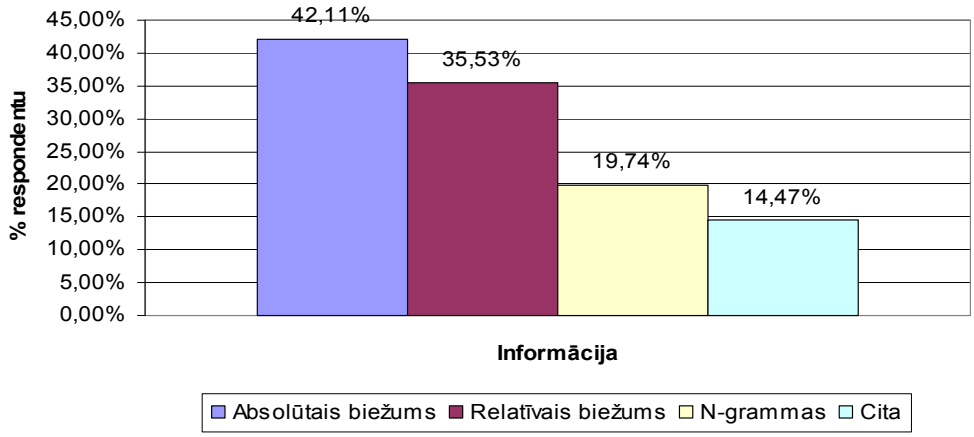
Respondentus visvairāk (52,63%) interesē vienkāršā meklēšana pēc konkrēta vārda, taču gandrīz puse respondentu (48,68%) norādījuši nepieciešamību izmantot sarežģīto meklēšanu (ne tikai pēc vārdformas, bet minot arī, piemēram, tipisku kontekstu). Vairāk nekā trešā daļa respondentu atzina, ka viņiem ir nepieciešama informācija par tipiskiem vārdu savienojumiem. Gandrīz trešdaļa respondentu uzskata, ka viņiem būtu nepieciešama cita informācija, piemēram, sinonīmu un antonīmu meklēšana, meklēšana pēc darbības vārdu pārvaldījuma, latviešu valodas vārdam atbilstoša svešvārda meklēšana, okazionālismu meklēšana konkrētas jomas tekstos. Tiek minēta arī inversās vārdnīcas nepieciešamība. Latviešu valodas gramatikas pētniekiem svarīga ir latviešu morfēmu produktivitāte, derivatīvās ligzdas – arī šādi pieprasījumi minēti kā citas iespējas. Citu starpā tika minētas meklēšana, izslēdzot konkrētu vārdu vai vārdu savienojumu un izvērstā meklēšana. Tiek pausta vēlme, lai meklējot tiek piedāvāti arī līdzīgi vārdi (izplūdušā atbilsme).

Lai nodrošinātu sarežģīto meklēšanu (piem., pēc pamatformas), ir jābūt morfoloģiski marķētam latviešu valodas korpusam. Savukārt sintaktiski marķētā

korpusā būtu iespējama meklēšana pēc pārvaldījuma. Tā kā respondenti ir minējuši arī semantisko meklēšanu (hiponīmi, hiperonīmi), tas būtu iespējams tikai nākotnē izveidojot korpusu ar semantisko marķējumu.

3.9. Kāda statistiskā informācija par valodu Jūs interesē?

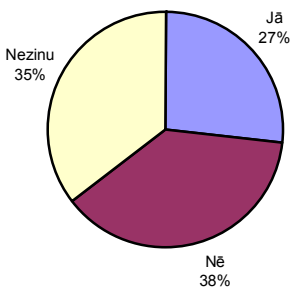
Absolūtais biežums, relatīvais biežums, n-grammas (burtu un vārdu savienojumi), cita.



3.13. diagramma – potenciālo lietotāju par valodu interesējošā statistiskā informācija.

Lai arī visvairāk lietotāju (42,11%) interesē valodas datu absolūtais biežums, diezgan ievērojams skaits respondentu (14,47%) ir norādījuši, ka vēlas uzzināt citu informāciju: vārdu savienojumu analīzi, konkrēta autora vārdlietojuma biežumu, vārda hiponīmus un hiperonīmus, atkāpes no valodas normām dažādos tekstos, jaunieviesto terminu lietošanas biežumu, abreviatūru biežumu dažādos tekstos, frazeoloģismu izmantošanas biežumu publicistikā un daiļliteratūrā, vārda tiešās un pārnestās nozīmes lietojuma biežumu.

3.10. Vai Jūs būtu ar mieru maksāt abonēšanas maksu par elektroniskajiem latviešu valodas resursiem?

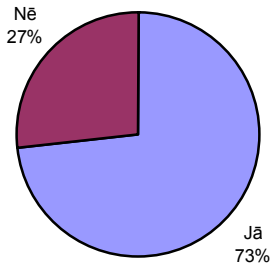


3.15. diagramma – respondentu attieksme pret abonēšanas maksu.

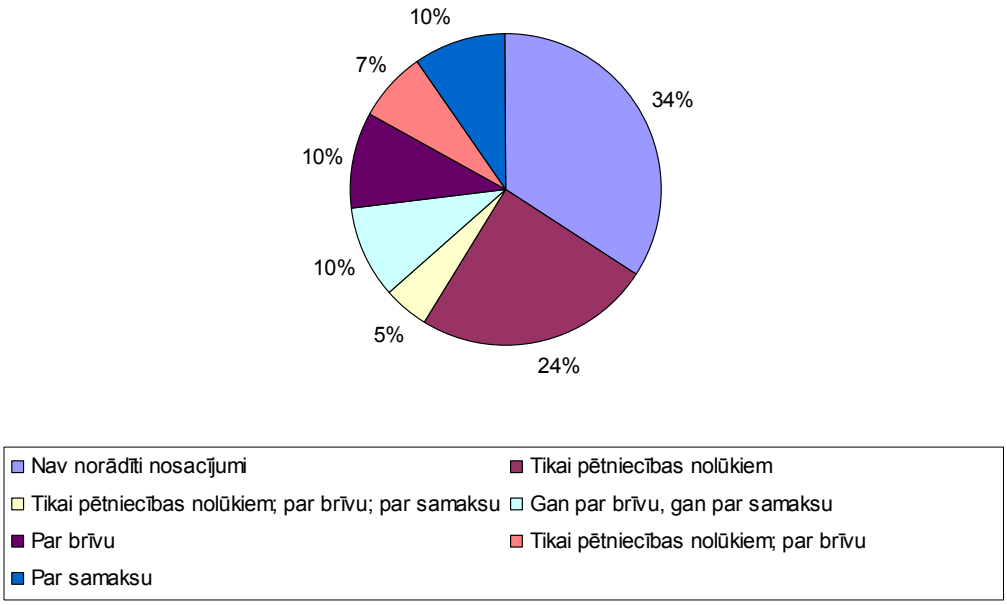
Šīs atbildes rāda, ka respondenti pieļauj iespēju maksāt abonentmaksu par noteiktas korpusa daļas izmantošanu atkarībā no lietojumu funkcionalitātes un korpusa satura. Nedaudz vairāk par trešdaļu respondentu ir noraidoši, bet gandrīz tikpat daudz respondentu vēl nav izlēmuši – tas būtu atkarīgs no piedāvājuma un abonentmaksas.

3.11. Vai Jūs būtu ar mieru piedāvāt savus elektroniskos resursus latviešu valodas korpusam?

Jā. / Nē. / Tikai pētniecības un mācību nolūkiem. Ja Jūs esat ar mieru piedāvāt savus elektroniskos resursus latviešu valodas korpusam, tad ar kādiem nosacījumiem?



3.16. diagramma – elektronisko resursu piedāvājums latviešu valodas korpusam.

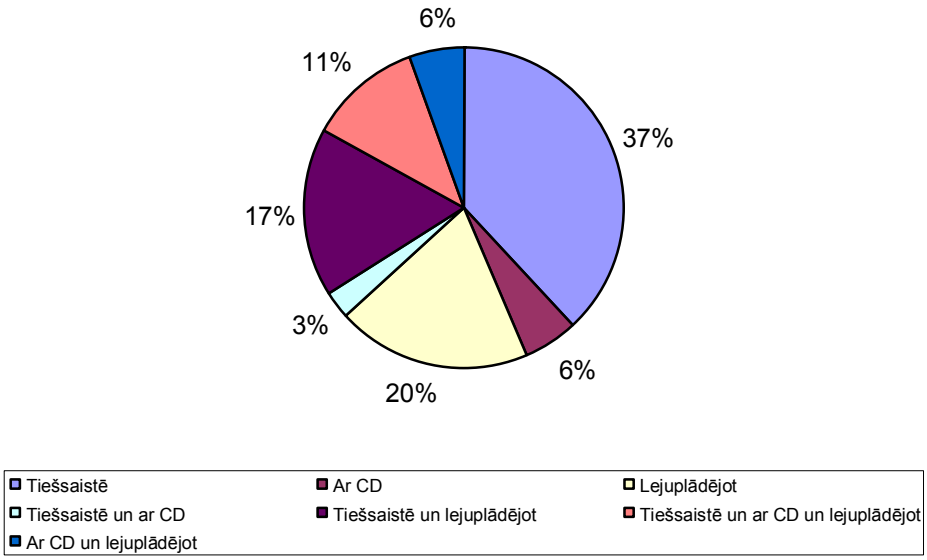


3.17. diagramma – nosacījumi, ar kādiem lietotāji piedāvātu savus elektroniskos resursus.

Lielākā daļa respondentu ir gatavi piedāvāt savus uzkrātos resursus latviešu valodas korpusam bez nosacījumi vai arī tikai pētniecības nolūkiem.

3.12. Kādā veidā Jums visērtāk izmantot elektroniskos resursus?

Tiešsaistē, ar CD, lejupielādējot?



3.18. diagramma – elektronisko resursu izmantošanas iespējas.

3.13. Secinājumi

Apkopojot aptaujas datus, var secināt, ka respondenti ir labi pazīstami ar latviešu valodas elektroniskajiem resursiem un konkrētiem ikdienā lietotiem programmrīkiem (pareizrakstības pārbaudītāju un automatizētām tulkošanas programmām), turklāt viņiem kā lietotājiem ir radusies nepieciešamība arī pēc tādas informācijas, ko vārdnīcas un tekstu krājumi, elektroniskas bibliotēkas nepiedāvā. Tā kā lielākā daļa respondentu (82%) strādā arī ar citu valodu datiem un no tiem 34% ir pazīstami ar citu valodu korpusu (un daudziem programmrīkiem), tad latviešu valodas korpuss kā jauns resursa veids būtu vērtīgs papildinājums.

Lielākoties respondenti izvēlas strādāt ar tekstiem bez marķējuma, jo tekstus ar papildinformāciju izmantotu tikai viena trešdaļa respondentu. Iespējams, ka jautājums tika pārprasts kā: *Vai jūs jau lietojat tekstus ar papildu informāciju*, nevis *lietotu tekstus ar papildu informāciju*. Tā kā respondenti ir ieinteresēti dažādas informācijas iegūšanā no valodas korpusa, tad loģisks ir secinājums, ka tādu datu izguve nav iespējam no nemarķēta valodas korpusa. Ja ir nepieciešamība uzzināt datus par pārvaldījumu, tad ir nepieciešams morfoloģiski un sintaktiski marķēts valodas korpuss un attiecīgs programmnodrošinājums.

Lielākā daļa respondentu, kas strādā ar citu valodu datiem, izmanto tos pārsvarā tulkošanā, tas liecina arī par nepieciešamību veidot divvalodu (daudzvalodu) paralēlos korpusus, piem., angļu-latviešu paralēlo korpusu, kuru varētu izmantot gan tulkošanas

procesā, gan tulku un tulkotāju apmācībā (sīkāk par divvalodu korpusa izmantošanu skatīt 4. nodaļā).

Respondenti norāda, ka visbiežāk strādā ar e-pasta vēstulēm, iespējams, ka savām vajadzībām ir uzkrājuši šādu resursu veidu, ko izmanto pētniecībā. Turklāt 73% respondentu ir gatavi piedāvāt savus resursus latviešu valodas korpusam (pētniecības nolūkiem un/vai apmaiņā pret iespēju izmantot citu valodas resursus).

Izdevniecību pārstāvji, kas iesūtījušas aptaujas, pagaidām nav gatavi piedāvāt savus resursus, bet koncepcijas autori iesaka latviešu valodas korpusa veidotājiem griezties pie lielākajām izdevniecībām un, sīkāk paskaidrojot korpusa satura sabalansētības svarīgumu un lietojumu iespējas, kā arī autortiesību jautājumu risinājumus, precizēt viņu nostāju šajā jautājumā.

Apmēram trešdaļa respondentu būtu ar mieru maksāt abonēšanas maksu par valodas korpusa izmantošanu, bet trešdaļa nemācēja atbildēt uz šo jautājumu, kas nozīmē, ka valodas tekstu korpuss vai atsevišķi tā lietojumi var tikt piedāvāti par maksu.

Vairums respondentu korpusu izmantotu tiešsaistē, bet daudzi arī vēlētos strādāt ar kompaktdisku un datus lejupielādējot, kas nozīmē, ka korpusa veidotājiem būs jādomā par datu izlases veida publicēšanu uz informācijas nesējiem.

4. Korpusa izmantojuma iespēju, pieejamības interesentiem un speciālistiem raksturojums

Aplūkojot citu valstu pieredzi un aptaujāto respondentu izteiktās vēlmes par latviešu valodas korpusu, koncepcijas autori piedāvā izveidot dažādu tipu korpusus, kuriem var būt dažādi lietojumi.

Koncepcijas autori piedāvā izveidot **vispārīgu latviešu valodas korpusu**, kurā būtu ievietoti sabalansētā veidā teksti, kas pārstāv mūsdienu latviešu valodu (no 20. gs. 80. gadu beigām). Papildus koncepcijas autori iesaka veidot arī **speciālos latviešu valodas korpusus** (piem., noteiktas izloksnes korpusu vai studentu (valodas apguvēju) korpusu).

20. gs. deviņdesmitajos gados radās nepieciešamība pēc jauniem datiem – divvalodu un daudzvalodu korpusiem. Atskatoties vēsturē, redzam, ka jaunajam korpusa tipam tika piedāvāti dažādi nosaukumi, piem., „divvalodu korpuss" [Gale, Church 1993], bet šis termins neieguva popularitāti un drīz tika aizstāts ar citiem ieteikumiem. Izskanēja aicinājums nošķirt korpusu, ko veido „oriģinālteksti un to tulkojumi", no tāda korpusa, ko veido „paralēli oriģinālteksti divās vai vairāk valodās" [Johansson, Hofland 1994], savukārt M. Beikere, viena no vadošajām translatoloģijas pētniecēm, aicināja ievērot atšķirības starp trīs dažādiem paralēlo korpusu veidiem:

(1) **paralēlo korpusu**, ko veido oriģinālteksti vienā valodā un to tulkojumi citā valodā;

(2) **daudzvalodu korpusiem**, kas ir vienvalodas korpusu kopa, kura veidota pēc vienotiem kritērijiem;

(3) **salīdzināmo korpusu**, ko veido paralēlie oriģinālteksti un tulkojumu teksti vienā valodā [Baker 1995].

Šodien visbiežāk runā par paralēlo korpusu, kurā ievietoti oriģinālteksti un to tulkojumi. Koncepcijas autori iesaka veidot angļu-latviešu paralēlo korpusu, ievietojot tajos gan oriģināltekstus, gan tulkojumus abās valodās, resp., tas būtu divvirzienu (*bi-directional*) korpuss.

4.1. Vispārīgā latviešu valodas korpusa izmantošanas iespējas

Vispārīgo latviešu valodas korpusu paredzēts plaši izmantot gan valodniecībā, gan datorzinātnēs, gan valodas mācīšanā, gan tulkotāju sagatavošanā, gan tulkošanas procesā. Minēsim dažas no valodniecības iespējām, atsevišķas jomas aprakstot sīkāk.

Valodas korpuss ir lietderīgs:

(1) mūsdienu valodas pētīšanai kopumā (skatot gan runāto, gan rakstīto valodu);

(2) leksikogrāfijā;

(3) veicot salīdzināmo valodas analīzi gan laika šķērsgriezumā (piem., diahroniskais korpuss), gan pēc funkcionālajiem stiliem (piem., zinātnisko tekstu valoda; juridisko tekstu valoda u. tml.), gan arī viena veida tekstu kopumā (piem., daiļliteratūras apakškorpuss, kurā ir gan oriģinālliteratūra, gan tulkotā literatūra);

(4) konkrēta autora vai stila pētīšanā, uzmanību pievēršot atšķirīgajam viena autora / stila valodā; jāņem vērā, ka nepieciešamības gadījumā var veidot viena autora valodas korpusu, bet nebūtu vēlams vispārīgajā korpusā ievietot daudz viena autora darbu, jo tādējādi iegūstam drīzāk subjektīvu, nevis objektīvu valodas ainu.

Valodas korpuss sniedz materiālu arī kultūrvēsturniekiem un literatūras kritiķiem, ļaujot plašākā diskursā skatīt gan atsevišķus jēdzienus, gan autorus (tai skaitā autorības jautājumu), gan noteikta laika iezīmes. Korpusu var izmantot arī lielāka konteksta pētījumos it kā nesaistītās jomās – tirgvedībā, politoloģijā, reklāmās un plašsaziņas līdzekļos, tiesu ekspertīzē u. c.

Dažus lietojumus apskatīsim sīkāk.

4.1.1. *Valodniecībā*

Leksikogrāfija. Viena no jomām, kur latviešu valodas korpuss būtu ļoti svarīgs, ir vārdnīcu veidošana, jo valodas korpuss sniedz kvantitatīvus, kā arī kvalitatīvi jaunus datus. Turklāt aizvien biežāk vārdnīcās tiek iestrādāti dati gan par runāto, gan par rakstīto valodu. Dažos vārdnīcu tipos ir svarīgi ņemt vērā un pievienot arī informāciju par vārda/vārdformas biežumu (piem., mācību vārdnīcās). Vārdnīcu veidošanā var izmantot gan sinhronisko, gan diahronisko valodas korpusu, ņemot vērā, kāda veida vārdnīcu iecerēts izveidot. Daži priekšlikumi:

(1) mūsdienu latviešu valodas skaidrojošās vārdnīcas sagatavošanā būtu ieteicams izmantot mūsdienu latviešu valodas korpusu (resp. sinhronisko korpusu), kas būtu vispārīgā korpusa sastāvdaļa;

(2) iespējas veidot jaunu latviešu valodas biežuma vārdnīcu, kā arī vārdformu biežuma vārdnīcu;

(3) ar konkordances palīdzību ir iespējams iegūt biežāk lietoto un tipiskāko vārdu savienojumu sarakstu, ko var iestrādāt skaidrojošās vārdnīcas šķirkļa sastāvā; vienlaicīgi iegūstam arī potenciālo stabilu vārdu savienojumu sarakstu (tai skaitā arī frazeoloģismus);

(4) izmantojot morfoloģiski anotētu korpusu, ar konkordances palīdzību mēs iegūstam priekšstatu arī par gramatikas modeļiem, piemēram, informāciju par darbības vārdu valenci, dažādu informāciju par biežumu u. tml.; šo

informāciju iestrādā gan vārdnīcas šķirklī, gan arī ir iespēja veidot tieši „gramatisko modeļu" vārdnīcu;

(5) strādājot ar diahronisko korpusu, iespējams veidot latviešu valodas vēsturisko vārdnīcu (piem., jau šobrīd, izmantojot latviešu valodas seno tekstu korpusu[32], ar LZP atbalstu ir uzsākts darbs pie latviešu valodas vēsturiskās vārdnīcas (16. – 18. gs.) koncepcijas izstrādes (projekta vadītājs prof. P. Vanags);

(6) latviešu valodas tēzaura izveide nav iedomājama bez lieliem latviešu valodas resursiem un, protams, arī diahroniskā korpusa;

(7) izmantojot latviešu valodas speciālo korpusu (piem., kādas izloksnes korpusu vai kāda funkcionālā stila korpusu), iespējams veidot jaunas vārdnīcas.

Pasaules prakse. Jau sākot ar pirmajiem angļu valodas korpusiem, kas tika veidoti dažādiem pētniecības mērķiem, galvenais akcents tika likts uz korpusa izmantošanu vārdnīcu veidošanā. Šādi korpusi tiek veidoti ne tikai no savāktiem tekstiem, bet arī no iepriekš uzkrātām kartotēkām un vārdnīcām, kas tika sagatavotas elektroniskā veidā (piem., *Longman Dictionary of Contemporary English* (LDOCE) izmantota kā korpuss). Šodien lielākajām angļu izdevniecībām, kas veido vārdnīcas, ir arī savi valodas resursi. Spilgtākie piemēri korpusa izmantošanai vārdnīcu veidošanā ir:

(1) *COBUILD* projekts, kas aizsākās 80. gados, šobrīd korpuss ir izaudzis par vienu no lielākajiem pasaulē (*Bank of English)*, izdevniecība *Harpers Collins* piedāvā internetā izmantot 56 milj. lielu paraugu[33];

(2) *Longman Corpus Network*[34], kas ietver sevī piecus lielus korpusus *(Longman Learners' Corpus, Longman Written American Corpus, Longman American Spoken Corpus, Spoken Corpus, Longman/Lancaster Corpus)*;

(3) *Cambridge International Corpus*[35] ar 700 milj. vārdlietojumu apjomu, kas ietver gan *Cancode* korpusu, gan *Cambridge Learner Corpus* u. c.

Publiski pieejama saite „Frāzes angļu valodā"[36], kur iespējams ievadīt interesējošo vārdu un iegūt gan konkordances, gan dažādus statistikas datus, izmantojot Britu nacionālā korpusa 2000. gada pasaules izdevumu. Šo informāciju var izmantot gan leksikas un semantikas pētījumos, gan iestrādāt vārdnīcas šķirkļos.

[32] http://www.ailab.lv/SENIE – skatīts 01.07.2005.
[33] http://www.collins.co.uk/Corpus/CorpusSearch.aspx – skatīts 18.07.2005., meklēšanas rezultāti netika sagaidīti.
[34] http://www.longman.com/dictionaries/corpus/lccont.html – skatīts 04.07.2005.
[35] http://www.cambridge.org/elt/corpus – skatīts 04.07.2005.
[36] http://pie.usna.edu – skatīts 04.07.2005.

Gramatikas pētījumi. Jau ar pirmo angļu valodas korpusu izveidi pētnieki apgalvoja, ka tagad ir iespējams pētīt un uzzināt „patiesos faktus" par angļu valodas gramatiku, tādējādi sevi pretstatot Homska tradīcijas piekritējiem, kas uzskatīja, ka ir jāpētī valodas kompetence, kas ir ideālā runātāja valodas zināšanas. Savukārt korpusa veidotāji uzsver, ka jāskata ir valodas performance – runātāja valodas izpildījums, kas lieliski atspoguļojas korpusā. Nenoliedzami, ka korpusa dati sniedz iespēju saskatīt valodas modeļus un „valodas dabu", par kuru neaizdomājamies, pirms nav kvantitatīvas liecības.

Mūsdienu latviešu valodas pētījumiem ir nepieciešama jauna mūsdienu latviešu valodas gramatika, kurā būtu ietverti dati gan par runāto, gan par rakstīto latviešu valodu. Lai veiktu pētījumus par latviešu valodas morfoloģiju, ieteicams izmantot morfoloģiski anotētu korpusu. Savukārt sintakses pētījumiem ieteicams sintaktiski anotēts korpuss.

Izmantojot morfoloģiski anotētu korpusu diahroniskā aspektā, iespējams papildināt zināšanas latviešu valodas vēsturē, valodas variantu izpētē, gramatiskās normas attīstībā.

Pasaules prakse. Pirmajos angļu valodas korpusos (Brauna un LOB korpusā) bija iekļauti tikai rakstītās valodas teksti. Savukārt Londonas-Lundas korpusā tika iekļauti arī runātās britu angļu valodas dati. Tādējādi bija iespējams *Comprehensive Grammar of the English Language* [Quirk et al. 1985] iestrādāt arī ziņas par runāto valodu, kā arī veikt salīdzinošu rakstītās un runātās valodas analīzi [sal. Hofland, Johansson 1982]. Viens no biežāk citētajiem darbiem angļu valodā, kas veltīts angļu valodas sintaksei un stilistikai, ir D. Baibera u. c. veiktais pētījums [Biber, Conrad, Reppen 1998], kurā izvirzīta doma, ka katram vārdam vai pat vārda nozīmei piemīt zināma sintaktiskās struktūras kopa, nerunājot jau par atsevišķa funkcionālā stila gramatikas īpatnībām. Korpuslingvistikas metode ļauj pamanīt leksikas un sintakses savstarpējo atkarību, pretēji viedoklim par gramatikas un leksikas stingru šķīrumu.

Atsevišķi ir jāmin sintaktiski anotētie korpusi, t. s. *Tree Bank*, piem., *Penn Treebank*[37] projekts, *Alpino Treebank*[38] projekts, bulgāru valodas sintaktiski anotētais korpuss *BulTreeBank*[39] (*HPSG-based Syntactic Treebank of Bulgarian*). Iegūtie dati sniedz plašu ieskatu attiecīgās valodas gramatiskajā struktūrā un šo struktūru iespējamajos variantos, kā arī sintaktiski anotētos korpusus plaši izmanto datorlingvistikā un dabīgās valodas apstrādē (informācijas ieguvē u. tml.).

[37] http://www.cis.upenn.edu/~treebank – skatīts 04.07.2005.
[38] http://odur.let.rug.nl/~vannoord/trees – skatīts 04.07.2005.
[39] http://www.bultreebank.org – skatīts 04.07.2005.

Prāgas anotētais korpuss *The Prague Dependency Treebank*[40] ir morfoloģiski, sintaktiski un semantiski anotēts korpuss, ko izmanto gan teorētiskos valodniecības pētījumos, gan mašīnmācīšanas metožu izmantošanā valodas analīzei un ģenerēšanas rīku izstrādē.

Vēsturiskā valodniecība. Vēsturiskajā valodniecībā atšķirībā no sinhronās valodniecības nav pieejami runātāji, bet plaši tiek izmantoti rakstītu materiālu dati. Agrāko kartotēku vietā aizvien biežāk tiek veidoti diahroniskie korpusi, kas vai nu aptver konkrētu periodu, vai arī sniedz valodas attīstības ainu plašākā periodā. Atsevišķos gadījumos diahroniskā korpusa pamatā var būt vārdnīca, kurā ir ietverta informācija par valodas vēsturi [sal. Ungārijas ZA izveidotais vēsturiskās vārdnīcas korpuss[41]].

Atšķirībā no sinhroniskā korpusa, kuram tiek izvirzīti vairāki izveides kritēriji, diahroniskajam korpusam netiek izvirzīta reprezentativitāte. Taču diahroniskajā korpusā parasti ir doti detalizēti metadati, kas ļauj atlasīt sev nepieciešamos avotus pēc vairākiem kritērijiem (piem., laiks, autors, teksta tips u. c.). Morfoloģiski vai sintaktiski anotējot korpusu, iegūstam ļoti vērtīgus datus. Diahronisko korpusu var izmantot:

(1) vēsturiskās gramatikas, valodas normēšanas vēstures pētījumos;

(2) var skatīt konkrēta laika perioda, kāda autora valodas īpatnības;

(3) iespējams pētīt valodas variantus visos valodas līmeņos (fonētikas, morfoloģijas, sintakses), jo tieši valodas varianti ir viena no agrākā laika tekstu pazīmēm;

(4) protams, ka diahroniskais korpuss ir nepieciešams jebkuras vēsturiskās valodas vārdnīcas izstrādei;

(5) var diahroniski pētīt viena teksta vairākus tulkojumus un veikt teksta analīzi, kā arī skatīt avotvalodas ietekmi pirmajos latviešu tekstos;

(6) arī viena autora valodu var pētīt diahroniskā aspektā.

2002. gada beigās tika izveidots latviešu valodas seno tekstu korpuss[42], kas tiek izmantots latviešu valodas vēsturiskās vārdnīcas (16. –18. gs.) izstrādes principu izveidei un paraugšķirkļu sagatavošanai, kā arī LU Filoloģijas fakultātes un Sanktpēterburgas Valsts universitātes studentu mācību procesā, lai iepazīstinātu ar latviešu valodas vēsturi un analizētu agrākos rakstu pieminekļus.

[40] http://ufal.mff.cuni.cz/pdt2.0 – skatīts 04.07.2005.
[41] http://www.nytud.hu/hhc – skatīts 18.07.2005.
[42] www.ailab.lv/SENIE – skatīts 18.07.2005.

Bez jau minētajām iespējām diahroniskajā korpusā, pakāpeniski atlasot avotus vai to fragmentus, ir iespējams veidot tā saukto parauga korpusu, kas varētu pārstāvēt noteikta gadsimta rakstītās valodas zināmu etalonu.

Lai izveidotu labu diahronisko korpusu, ir jāņem vērā vairāki nosacījumi:

(1) hronoloģija – korpusā jābūt pārstāvētiem vairākiem laika vai vēstures periodiem;

(2) reģionālās īpatnības – jāpievērš uzmanība tam, lai korpusā tiktu pārstāvētas valodas dažādās reģionālās īpatnības (lai arī galvenie latviešu valodas rakstu pieminekļi pārstāv vidus dialektu, ir arī augšzemnieku rakstu pieminekļi, kurus var un vajag pievienot latviešu valodas seno tekstu korpusam);

(3) sociolingvistiskais aspekts – latviešu valodas rakstu vēsturē 16. un 17. gs. dominē baltvācu mācītāju rakstītie teksti, tomēr tas tāpat ļauj tos pēc vajadzības grupēt sīkāk (piem., tie, kas ir dzimuši Latvijā, tie, kas ir ieceļojuši šeit pēc studijām utt.);

(4) stilu (vai žanru) daudzveidība – lai arī 16. un 17. gs. dominē garīga satura teksti ar nedaudziem juridiskiem tekstiem un valodniecisko literatūru, 18. gs. raksturīgs plašs stilu klāsts, kas sniedz iespēju analizēt latviešu valodu vēsturiskā aspektā arī pēc valodas funkcionālajiem stiliem.

Seno tekstu korpuss ļautu novērtēt seno tekstu lomu latviešu valodas vēstures pētījumos vispār, jo trūkst latviešu valodas attīstības pētījumu. Tā kā līdz salīdzinoši nesenam laikam latviešu valodniecībā bija novērojama ļoti kritiska attieksme pret senajiem tekstiem kā valodas vēstures avotu (nešķirot kultūrvēsturisko un filoloģisko aspektu), diahroniskais korpuss ļautu objektīvi skatīt un izvērtēt agrākos rakstu pieminekļus.

Pasaules prakse. Pasaulē pazīstamākais angļu valodas diahroniskais korpuss ir Helsinku korpuss[43] (*Helsinki Corpus of English Texts*), kuru veido tekstu izlases no 700. –1700. gadam, kā arī dialektu korpusa daļa, kuru veido intervijas ar lauku iedzīvotājiem 1970. gados. Šī korpusa uzdevums ir veicināt un popularizēt diahroniskos pētījumus, kā arī darīt pieejamos agrāk publicētos tekstus plašākai sabiedrībai [Rissanen 1992].

Seno tekstu korpuss, kas ir daļa no lielā Helsinku korpusa, šobrīd ir pabeigts un pieejams akadēmiskiem mērķiem. Šeit ir apmēram 1,5 miljonu vārdlietojumi, un tajā ir ap 400 tekstu paraugiem. Īsākie teksti ir iekļauti pilnā apjomā, bet no lielākiem tekstiem tika izvēlēti 2 500 – 20 000 vārdlietojumu lieli fragmenti.

[43] http://khnt.hit.uib.no/icame/manuals/HC/index.htm – skatīts 18.07. 2005.

Helsinku korpusa veidotāji pievērsa uzmanību tam, lai viņu korpuss būtu plaši izmantojams un atbilstu dažādiem parametriem, tāpēc attiecīgie teksti un fragmenti ir izvēlēti pēc rūpīgas sociolingvistiskās analīzes. Izmantojot COCOA formātā rakstītus kodus (tos izmanto Oksfordas konkordances programma (saīsināti – OCP)), uzreiz tiek sniegta informācija par tekstu, tā žanru un autoru:

(1) no kāda teksta ir ņemts fragments (ja tas ir fragments);

(2) vai teksts ir privāts vai publisks;

(3) laiks, kad teksts ir sarakstīts;

(4) ja tā ir vēstule, vai sūtītājs un adresāts ir vienādā sociālā stāvoklī;

(5) autora dzimums;

(6) autora vecums.

Šāds marķējums korpusa lietotājam sniedz galveno informāciju par tekstu vai tā fragmentu. Dažas atpazīšanas programmas, piem., OCP, izmanto šādu kodēto informāciju, lai pēc tam veiktu meklēšanu tikai pēc noteiktajiem atlases kritērijiem. Parametru kodēšana ļauj izmantot korpusu divējādi:

(1) skatīt visus vārda lietojumus visā korpusā un tad meklēt atšķirības pēc dažādiem parametriem (piem., dialekts, teksta žanrs, autors);

(2) izvēloties noteiktus parametrus, var atlasīt tikai konkrēta perioda vai žanra piemērus.

Kā parauga korpusa piemēru var minēt ARCHER korpusu (*A Representative Corpus of Historical English Registers*), kurā ir 1,7 miljonu vārdlietojumu no 1037 tekstu paraugiem. Šis korpuss pārstāv balansētu septiņu rakstītās valodas stilu (dienasgrāmatas, vēstules, daiļliteratūra, ziņas, zinātne u. c.) un trīs runā balstītas valodas stilu (daiļdarbu sarunas, dramaturģijas un spredikļi) paraugus [Biber, Finegan, Atkinson 1994].

Pieejami arī noteiktu tekstu stilu (piem., angļu valodas dzejas korpuss *Corpus of Middle English Prose and Verse*[44] vai Londonā iznākušo avīžu korpuss *Zurich English Newspaper Corpus (ZEN),* kurā iekļauti no 1671. līdz 1791. gadam iznākušo avīžu teksti [sal. Fries, Schneider 2000]); teksta veidu (piem., angļu valodas dialogu korpuss (CED), kurā ir iekļauti dialogu teksti, kas datēti ar 1560.–1760. gadu) diahroniskie korpusi. Arī daudzu nacionālo korpusu sastāvā ir diahroniskie korpusi (sk. piem., Čehu nacionālo korpusu[45]).

Dialektu pētījumi. Izmantojot speciālo korpusu, kurā ir iekļauta gan runātā, gan rakstītā valoda, iespējams veikt dažāda līmeņa pētījumus par kādu konkrētu dialektu.

[44] http://www.hti.umich.edu/c/cme/about.html – skatīts 29.07.2005.
[45] http://ucnk.ff.cuni.cz – skatīts 29.07.2005.

Latviešu valodas dialektiem ir gan rakstu pieminekļi, gan runātās valodas dati, kas savākti dažādu ekspedīciju laikā vairākās augstskolās un pētniecības iestādēs. Tādējādi varētu izveidot latviešu valodas izlokšņu korpusu ar detalizētu metainformāciju, lai varētu veikt precīzu un izvērstu meklēšanu. Sasaistot skaņu ierakstus ar atšifrēto tekstu, šādu korpusu varētu izmantot vispārīgiem dialektu pētījumiem. Viens no problēmjautājumiem, ar ko sastapsies šāda korpusa veidotāji, ir fonētiskās transkripcijas izvēle, jo Latvijā līdz šim dialektu tekstu atšifrēšanā nav ieviests vienots pieraksts.

Pasaules prakse. Speciālā *Survey of English Dialects*[46] korpusa pamatā ir Anglijas lauku apvidū ierakstītas intervijas. Laikā no 1948. –1961. gadam tika ierakstītas aptuveni 60 stundu sarunas, kurā sešdesmitgadīgi cilvēki stāstīja savas atmiņas par ģimeni, darbiem. Runātās valodas korpusā kopā ir 800 000 vārdu. Ieraksti tika transkribēti, skaņu faili sasaitīti ar tekstu atšifrējumiem. Turklāt, lai šādu korpusu varētu izmantot pēc iespējas pilnvērtīgāk, tajā ir veikta vārdšķiru noteikšana un marķēšana. Korpuss ir nenoliedzami vērtīgs materiāls gan dialektologiem, gan valodas vēstures pētniekiem, gan kultūrvēsturniekiem. Šī korpusa CD publicēja *Routledge* izdevniecība Londonā.

Lielākais angļu valodas dialektu korpuss izveidots Freiburgā *The Freiburg English Dialect Corpus (FRED)*[47]. Korpusa veidotāju mērķis bija uzkrāt datus, lai varētu veikt pētījumus un izdarīt vispārinājumus par dažādu dialektu gramatiku gan kvantitatīvajā, gan kvalitatīvajā līmenī. Ierakstot materiālu no lielākajiem britu angļu valodas dialekta reģioniem un apkopojot materiālu no dzīvesstāstu projektiem, korpusa apjoms sasniedzis 2,4 miljonus (te iekļauti 370 teksti no 9 apgabaliem. Runātāji, kuru teksti iekļauti korpusā, dzimuši laika posmā no 1890.–1920.gadam, un sarunas tika ierakstītas 20. gs. septiņdesmitajos un astoņdesmitajos gados. Šobrīd korpuss tiek izmantots angļu valodas gramatikas un valodas variantu pētījumos, vienlaicīgi veidojot arī sastatāmo gramatiku [Kortmann et al 2005].

Igaunijā ir izveidots igauņu valodas dialektu korpuss[48].

Psiholingvistika un sociolingvistika. Veidojot speciālos korpusus, iespējams iegūt jaunus datus dažādās psiholingvistikas un sociolingvistikas pētījumu jomās. Lai pārbaudītu hipotēzes, kā saistīta smadzeņu un valodas darbība, nepieciešami mentālo procesu mērījumi. Šādus datus var iegūt, ierakstot runu un vienlaicīgi fiksējot arī attēlu. Savukārt, vispārīgo vai speciālo korpusu sīkāk analizējot pēc metadatiem (piem.,

[46] http://www.yorksj.ac.uk/dialect/SED.htm – skatīts 29.07.2005.
[47] http://www.anglistik.uni-freiburg.de/institut/lskortmann/ FRED – skatīts 29.07.2005.
[48] http://www.murre.ut.ee – skatīts 29.07.2005.

vecuma, dzimuma, nodarbošanās, dzīvesvietas, izglītības līmeņa u. tml.), iegūstam informāciju par tipiskām parādībām noteiktas sociālās grupas valodā.

Daži korpusa lietojumi:

(1) izveidojot, piem., bērnu valodas korpusu, var izsekot līdzi valodas apgūšanas problēmām; tas lieti noder ne tikai psiholingvistikā, bet arī datorlingvistikā automatizētu sistēmu modelēšanā;

(2) veidojot speciālu noteiktas vecuma grupas (piem., jauniešu (pusaudžu)) valodas korpusu, iespējams izsekot noteikta slenga attīstībai, kā arī fiksēt ekstralingvistisko faktoru ietekmi uz runāto un rakstīto valodu;

(3) izmantojot latviešu valodas runātās valodas korpusu, iespējams pētīt runas kļūdas dzīvā sarunvalodā un tās klasificēt;

(4) iespēja pētīt valodas pataloģijas gadījumus.

Pasaules pieredze. Viens no pazīstamākajiem bērnu valodas uzkrāšanas un analīzes projektiem ir *Child Language Data Exchange System* korpuss (CHILDES)[49], kurā ir aptuveni 20 milj. vārdlietojumu dati 20 pasaules valodās. Datus, kas iegūti šajos pētījumus, piemēro ne tikai pirmās, bet arī otrās valodas apguves problēmu risināšanai.

Lai izpētītu bērnu valodas sintaksi, 1978.–1984. gadā tika uzkrāts *Polytechnic of Wales (PoW)*[50] korpuss. Pierakstot 6–12 gadu vecu bērnu valodu, tika izveidots 65 000 vārdlietojumu korpuss, kas tika sintaktiski izanalizēts.

Rakstītās bērnu valodas korpusā *Lancaster Corpus of Children's Project Writing* (LCCPW)[51] fiksēti 9–11 gadus vecu bērnu ar roku rakstīti sacerējumi. Apvienojot vizuālo un tekstuālo informāciju, kā arī pievienojot vārdšķiru marķējumu, šis korpuss ļauj iegūt dziļāku priekšstatu par bērnu valodas attīstību.

Slavenāko jauniešu valodas korpusu vidū minams Bergenas COLT korpuss *Corpus Of London Teenage Language*[52].

4.1.2. *Valodas mācīšanā*

Gan vispārīgo, gan arī speciālo latviešu valodas korpusu var izmantot latviešu valodas apguvē. Korpusa dati sniedz iespēju studentiem iepazīt autentiskus runātās un rakstītās valodas tekstus, iepazīt valodas diskursu. Korpuss ļauj likt uzsvaru uz biežāk lietoto leksiku un atspoguļo leksisko vienību izvēli. Vienlaicīgi korpusa izmantošana ietekmē arī valodas mācīšanas metodoloģiju.

[49] http://childes.psy.cmu.edu – skatīts 29.07.2005.
[50] http://khnt.hit.uib.no/icame/manuals/pow.htm – skatīts 29.07.2005.
[51] http://bowland-files.lancs.ac.uk/lever/index.htm – skatīts 29.07.2005.
[52] http://www.hf.uib.no/i/Engelsk/COLT/index.html – skatīts 29.07.2005.

Izmantojot valodas korpusu, uzsvars no vārdformu vai biežuma indeksa analīzes tiek pārlikts uz konkordances rezultātu izpēti. Pirmkārt, konkordanci var izmantot **leksikas apguvē**:

(1) var skolēnus (studentus) aicināt no konteksta uzzināt vārda nozīmi, uzrakstot pašiem savu definīciju un tad salīdzināt ar vārdnīcas šķirkļa informāciju, secinot, kas ir parādīts, bet kas netiek minēts vārdnīcā;

(2) var skatīt atsevišķu vārdu stilistisko nokrāsu;

(3) iepriekš sagatavojot konkordances rindiņas un ļaujot izmantot vārdnīcas, veikt salīdzinājumu, cik un kādas vārda nozīmes, kas sastopamas valodas korpusā, ir fiksētas un cik lielā mērā atspoguļotas vārdnīcā;

(4) izmantojot konkordances rindiņas ar izlaistu atslēgas vārdu, var skolēnus (studentus) aicināt pēc konteksta uzminēt, kāds ir izlaistais vārds;

(5) var apgūt stabilos vārdu savienojumus un frazeoloģismus;

(6) skatot sinonīmisku vārdu konkordances rindiņas (piem., *liels* un *neliels*; *mazs*, *sīks* un *neliels*), likt noskaidrot, ar kādiem vārdiem tie sastopami, kādas kombinācijas ir vienīgās iespējamās, bet kuros gadījumos šos īpašības vārdus var mainīt vietām.

Tāpat ir iespējams, izvēloties konkrēta autora vai stila tekstus, apgūt speciālo leksiku, piem., izvēloties folkloras tekstu apakškorpusu, var iepazīties ar vārdiem, kas ir laika gaitā mainījuši savu nozīmi, kā arī arhaismiem, historismiem, apvidvārdiem un barbarismiem. Tādējādi tas kalpotu gan par pamatu folklorā sastaptās leksikas analīzei, gan nākotnē varētu domāt par atsevišķas vārdnīcas izveidi. Ideāli, ja var strādāt ar atsevišķiem apakškorpusiem, piem., daiļliteratūras korpusu, juridisko tekstu korpusu, tad var piedāvāt, piem., poētiskās, zinātniskās, juridiskās valodas pētījumus.

Turklāt var domāt par korpusā balstītas analīzes iekļaušanu arī divvalodu vārdnīcās, jo bieži vien studenti, lasīdami mājās (folkloras vai latviešu literatūras klasiķu) tekstus, ar grūtībām tiek galā, jo šāda leksika lielākoties netiek iekļauta divvalodu vārdnīcās.

Otrkārt, latviešu valodas korpusu var izmantot **gramatikas apguvei**, jo korpusā sastopam dzīvās valodas, nevis pašu izdomātus piemērus. Skolēniem (studentiem) var sagatavot uzdevumus, kurā pēc konkordances rindiņām ir jānoskaidro, piem., prievārdu lietošana, darbības vārdu valence, vārdu savienojumu lietojums u. tml. Strādājot ar morfoloģiski anotētu korpusu, var izmantot izvērsto meklēšanu, tādējādi skolēns (students) labāk apgūst valodas gramatiskos modeļus un informāciju par vārdu saistāmību.

Atsevišķs korpusa veids ir **studentu (valodas apguvēja) korpuss**, kurā iekļauti cittautiešu rakstītie sacerējumi. Šādu korpusu varētu veidot, gan savācot no Latvijas

skolām cittautiešu skolēnu sacerējumus, kas rakstīti latviešu valodas stundās, gan arī cittautu studentu darbus, ko raksta ārpus Latvijas.

Pirmkārt, šāds korpuss ir īpaši noderīgs kļūdu analīzes veikšanai, lai noskaidrotu, kas īpaši jāņem vērā, mācot latviešu valodu. Šāds studentu korpuss, to attiecīgi sagatavojot, ļauj iegūt datus arī par tipiskiem kļūdu gadījumiem pēc vecuma, izglītības, tautības u. c. segmentiem, nodrošinot pamatu sociolingvistiskam pētījumam.

Otrkārt, šāds korpuss sniedz iespēju veikt arī sastatāmo analīzi, salīdzinot studentu un dzimtās valodas runātāju valodas līmeni. Šādus datus var izmantot arī nodarbībās, lai valodas studenti veiksmīgāk apgūtu sarežģīto gramatikas un leksikas materiālu.

Treškārt, šie dati noderētu jaunu latviešu valodas mācību grāmatu sagatavošanā.

Ceturtkārt, var veidot dažādus datorizētus mācību līdzekļus, kam piesaistīts arī valodas korpuss.

Pasaules prakse. Lielākais angļu valodas kā svešvalodas korpuss ir *International Corpus of Learners' English (ICLE)*[53]. Jāpiemin arī Longmana studentu korpuss[54], kas sasniedzis jau 10 miljonu vārdlietojumu.

Vienlaikus populāri un lietderīgi ir īpaši speciālās valodas korpusi (piem., Tamperē uzsāktais ELFA korpuss *English as a Lingua Franca in Academic Settings*[55]) kas izveidots, lai pētītu akadēmiskās valodas (lekciju, semināru u. tml.) īpatnības dažādu kultūru diskursā).

Tiek apkopoti ne tikai rakstītās valodas apguvēju runas dati, bet tiek veidoti arī runātās valodas korpusi (piem., *ISLE* (*Interactive Spoken Language Education*) korpuss[56], kurā apkopoti angļu valodas runātāju dati, kam tā nav dzimtā valoda).

Internetā ir pieejama *WordPilot* saite[57], tās mērķis ir palīdzēt valodas studentam pilnveidot vārdu krājumu, atrast pareizo vārdu, parādot piemērus no korpusa, sasaistot to ar vārdnīcas definīcijām, norādot iespējamos vārdu savienojumus. Tiek nodrošināta iespēja arī noklausīties izvēlēto tekstu.

Arī internetā pieejamā angļu valodas gramatika[58] ir saistīta un balstīta 1 miljonu lielā gramatiski anotētā britu runātās un rakstītās angļu valodas korpusa britu daļā ICE-GB[59].

[53] http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cecl-Projects/Icle/icle.htm – skatīts 29.07.2005.
[54] http://www.longman-elt.com/dictionaries/corpus/lclearn.html – skatīts 29.07.2005.
[55] http://www.uta.fi/laitokset/kielet/engf/research/elfa/index.htm – skatīts 29.07.2005.
[56] http://nats-www.informatik.uni-hamburg.de/~isle – skatīts 29.07.2005.
[57] http://www.compulang.com – skatīts 29.07.2005.
[58] http://www.ucl.ac.uk/internet-grammar – skatīts 29.07.2005.
[59] http://www.ucl.ac.uk/english-usage/ice-gb – skatīts 29.07.2005.

4.1.3. *Tulkošanā un tulkošanas studijās*

Tulku un tulkotāju ikdienas vajadzībām noder gan vispārīgais korpuss, gan speciālais (kādas nozares) korpuss, lai apzinātu terminoloģiju. Korpusu izmanto arī tulkošanas procesā, piem., konsultējoties, meklējot konkrētu vārdu savienojumu. Korpusa datus var iestrādāt daļēji automatizētas tulkošanas, arī uz piemēriem balstītas tulkošanas un statistiskās mašīntulkošanas sistēmās.

Papildus vispārīgais latviešu valodas korpuss būtu izmantojams arī tulkošanas studijās:

(1) skatot tulkoto tekstu avotvalodas ietekmi uz latviešu valodu;

(2) veicot tulkotāju izvēles analīzi;

(3) analizējot dažādus valodas stilus un apgūstot speciālo terminoloģiju.

Pasaules prakse. Mančestras universitātes Tulkošanas un starpkultūru pētījumu centrā Anglijā M. Beikeres vadībā ir izveidots angļu valodas tulkojumu korpuss – *Translation English Corpus (TEC)*[60]. Tas ir mūsdienu angļu valodas korpuss, kuru veido tulkojumi 10 milj. vārdlietojumu apjomā no daudzām avotvalodām angļu valodā. Šeit ir pārstāvēta daiļliteratūra, ziņas, žurnāli, ko lidsabiedrības izsniedz lidojuma laikā, kā arī biogrāfiju teksti. Tas nav divvalodu korpuss, bet gan speciāls vienvalodas korpuss, kas paredzēts tulkojumu valodas analīzei un tulkotāju sagatavošanai.

4.1.4. *Dabīgās valodas apstrādē*

Valodas korpusu datus tieši vai pastarpināti izmanto gan programmnodrošinājuma izstrādei (kā zināšanu avotus), gan to testēšanai. Zināšanas, ko iegūst no korpusa, izmanto dabīgās valodas varbūtiskā modelēšanā, bet statistiskos datus plaši izmanto valodas inženierijā. Te noder gan nemarķēti teksti, gan morfoloģiski, sintaktiski un semantiski marķēti korpusi. Daži no iespējamiem latviešu valodas korpusa lietojumiem:

(1) lai nodrošinātu pilnīgus datus par latviešu valodas statistisko informāciju un vārdu savienojumu (bigrammu, trigrammu) analīzi, ir nepieciešams sabalansēts vispārīgs latviešu valodas korpuss, kura analīzes dati izmantojami mašīntulkošanas sistēmās;

(2) pareizrakstības pārbaudītāja testēšanai, leksikona papildināšanai ar jauniem vārdiem, kas sastopami speciālajā vai vispārīgajā valodas korpusā;

(3) gramatikas pārbaudītāja izstrādē un pilnveidē;

(4) gramatiskās, leksikosintaktiskās informācijas ieguvē, kas noder ontoloģiju automātiskai būvēšanai, arī terminu izguvē;

(5) runas korpusa datu apstrāde runas sintēzes un analīzes rīku izveidošanai;

(6) valodneatkarīgu programmrīku apmācīšanai;

[60] http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm – skatīts 14.07.2005.

(7) jautājuma – atbilžu sistēmu izstrādē;

(8) informācijas izguvē tīmeklī;

(9) mākslīgā intelekta sistēmās.

4.2. Divvalodu (daudzvalodu) korpusu izmantošanas iespējas

Paralēlos korpusus, kas visbiežāk ir sastatīti teikuma līmenī (vai arī vārda līmenī), izmanto dažādās nozarēs. Daudzas no iespējām dublējas ar vienvalodas korpusa izmantošanu. Jau minēts, ka koncepcijas autori iesaka izveidot angļu-latviešu valodas paralēlo korpusu, kas ir sastatīts teikuma līmenī. Citi potenciālie valodu pāri būtu: latviešu-franču, latviešu-krievu paralēlais korpuss.

Sastatāmā valodniecība. Paralēlo korpusu dati ļautu pētniekiem Latvijā un ārpus tās veikt visu līmeņu valodas analīzi, izmantojot lielāka apjoma un daudzveidīgākus datus. Johansons un Hoflands [1994] uzsver novitāti, ko valodniecībā ienesa paralēlie korpusi: „divvalodu korpusi sniedz liecību par līdzīgo un atšķirīgo divās valodās. Tie ļauj veikt tekstā balstītu sastatāmo analīzi, kamēr tradicionālo sastatāmās valodniecības pētījumu uzmanība bieži vērsta uz abstraktas valodas sistēmas salīdzinājumu vai arī uz valodas sistēmas daļu salīdzinājumu, nepiesaistot to īstiem tekstiem" [Johansson, Hofland 1994:25].

Pasaules prakse. Viens no populārākajiem paralēlajiem korpusiem ir angļu-norvēģu paralēlais korpuss[61], jāmin arī tā paplašinājums *Oslo Multilingual Corpus* (OMC)[62], kā arī CONSTRA korpuss *Contrastive Studies in a Translation Perspective*[63] un Heimnicā veidotais angļu-vācu tulkojumu korpuss[64].

Varam minēt tikai dažus pētījumus, kuros izmantoti paralēlo korpusu dati: Vikbergs [Wikberg 1996] skata jautājumu teikumus angļu un norvēģu valodā; Altenbergs [Altenberg 1998] pētī saikļus un to, kā sākas teikumi angļu un zviedru valodā; Vibergs [Viberg 1998] aplūko daudznozīmīgus vārdus *run* un *put* angļu un zviedru valodā; Ebelings [Ebeling 1998] analizē angļu valodas kosntrukcijas ar *there* un tās tuvāko norvēģu valodas atbilsmi – konstrukcijas ar *det*; Johansons, pievēršot uzmanību angļu un zviedru valodas teikumu sākumam, uzsver, ka „dažu valodniecības parādību salīdzināšanu divās vai vairāk valodās vislabāk veikt ar tekstā balstītu pieeju" [Johansson 1996:37].

[61] http://www.hf.uio.no/iba/prosjekt – skatīts 14.07.2005.
[62] http://www.hf.uio.no/iba/OMC/English/index_e.html – skatīts 14.07.2005.
[63] http://www.hum.gu.se/~engadel/constra/index.html – skatīts 14.07.2005.
[64] http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/transcorpus/index.htm – skatīts 14.07.2005.

Lai veiktu juridisko tekstu sastatāmo analīzi, ir izveidots *BoLC Italian - English Comparable Corpus (Bononia Legal Corpus)*[65].

Leksikoloģijas pētījumi un leksikogrāfija. Leksikogrāfu galvenais darba rīks šodien ir valodas korpusi un esošās vārdnīcas. Leksikas pētījumu rezultātus, kurus iegūst no paralēlo korpusu datiem, var iestrādāt divvalodu vārdnīcās vai izmantot tos leksikoloģijas sastatāmajā analīzē. Jebkura leksikogrāfijas darbība saistīta ar milzīgu datu apstrādi. Paralēlo korpusu izmantošana atvieglos šādu datu sagatavošanu un, iespējams, veicinās jaunu divvalodas vārdnīcu tipu rašanos.

Paralēlo korpusu dati sniedz iespēju analizēt avotvalodas un mērķvalodas šķirkļa vārdus, to semantiskos tīklus, kā arī sniedz lielisku iespēju konstatēt, kurš no tulkojuma ekvivalentiem ir konteksta nosacīts, bet kurš ne. Paralēlā korpusa dati nodrošina virkni tulkošanas ekvivalentu, kurus diez vai var atrast iepriekšējās vārdnīcās – tas saistīts arī ar korpusa saturu. No paralēlajiem korpusiem ir iespējams iegūt ne tikai tulkojuma ekvivalentus vārdu, bet arī vārdu savienojumu līmenī, kurus var iekļaut vārdnīcā.

Pasaules prakse. Var minēt, piemēram, angļu un vācu valodas leksikas pētījumus (angļu valodas prievārdu *with* un tā tulkošanas ekvivalenti vācu valodā sk. [Schmied 1998; Schmied, Fink 2000]). Diezgan populāri ir salīdzināt korpusa liecības ar informāciju, ko sniedz divvalodu franču-angļu vārdnīcas [Dickens, Salkie 1996]. Veicot korpusa datu analīzi, parasti tiek secināts, ka „vienkārši vārdšķirās balstīti tulkojuma ekvivalenti, ko atrodam divvalodu vārdnīcās, neapmierina [Schmied 1998:271].

Viens no veidiem, kā piedāvāt izvēlēties strukturāli jaunus tulkojuma ekvivalentus, ir vārdnīcā iekļaut vairāk informācijas par gramatiku. Tiek demonstrēts mēģinājums, kā divvalodu vārdnīcas šķirklī sniegt pietiekamu gramatisko informāciju [Schmied 1998; Schmied, Fink 2000].

Tulkošanas studijas. Paralēlie korpusi tiek izmantoti tulkošanas atmiņu veidošanā un ikdienas tulkotāju darbā. Paralēlos un arī speciālos nozaru korpusus izmanto arī tulku un tulkotāju apmācībai. Paralēlie korpusi sniedz plašu semantisko informāciju, kas ne vienmēr ir atspoguļota vārdnīcās, jo šeit rodam kontekstā balstītus tulkošanas ekvivalentus. Turklāt paralēlo korpusu gadījumā var veidot apakškorpusus un skatīt tikai attiecīgo valodu tulkojumu korpusus, tādējādi analizējot tikai tos gan praktiskā, gan teorētiskā līmenī.

Pasaules prakse. Lai arī lielākā daļa pētījumu, kas veikti par korpusa izmantošanu translatoloģijā, veltīti rezultātu apskatam, M. Beikere izvirza svarīgu jautājumu par

[65] http://www.cilta.unibo.it/SITOBOLC_ITA.htm – skatīts 14.07.2005.

tulkojuma vietu korpuslingvistikā. Viņa ir izanalizējusi cēloņus, kāpēc korpuslingvistikai nav izdevies līdz šim ietekmēt tulkošanas studijas, kā pašu galveno minot faktu, ka korpuslingvisti tradicionāli izvairās no tulkojumiem un neiekļauj tos korpusā, jo tulkojumi nepietiekami reprezentatīvi atspoguļojot pētāmo valodu [Baker 1999].

Viena no metodoloģijas priekšrocībām, izmantojot tulkotos tekstus sastatāmā analīzē, ir tā, ka rezultātus, kas iegūti no oriģināltekstu salīdzināšanas, var piemērot tulkojumiem. Zviedru valodnieks M. Gellerstams [Gellerstam 1986], novērojot avotvalodas ietekmi uz tulkojuma valodu, ieviesa īpašu *translationese* jēdzienu. Turpinot analizēt šo parādību, norvēģu zinātnieki Johansons un Hoflands raksturo to kā „novirzi tulkojuma tekstos, ko izraisījusi avotvaloda" [Johansson, Hofland 1994:26]. Vācu pētnieki Šmīds un Šeflere [Schmied, Schäffler 1996] izmantoja šo pieeju, lai aprakstītu atšķirības angļu un vācu valodas tekstos, un sniedza detalizētu strukturālu un tipoloģisku atšķirību raksturojumu avota un mērķvalodas tekstos. Kings [King 1997] ziņo par daudzvalodu paralēlās konkordances izmantošanas projektu, kurā ietvertas vairākas Eiropas valodas, un uzsver trīs dažādas iespējas, kā daudzvalodu paralēlo korpusu var izmantot tulkošanas studijās:

(1) „pētot tulkotāja uzvedību sistemātiskā un principāli vispusīgā veidā";

(2) „salīdzinot tulkotāja izvēli ar divvalodu vārdnīcas piedāvājumu";

(3) „pārbaudot viedokļus, kas izteikti tulkošanas teorētiskā skatījumā" [King 1997:397].

Lai divvalodu (daudzvalodu) korpusu varētu izmantot tieši tulkošanas studijās, tam (tāpat kā jebkuram vienvalodas korpusam) ir jāatbilst zināmiem kritērijiem. Vācu pētniece Ulriha min šādus kritērijus:

(1) korpusā ir jāiekļauj pilni teksti (ne tekstu izlases); valodas un tulkojuma modeļus var pētīt tikai pilna apjoma tekstos;

(2) plaša diapazona tekstu iekļaušana korpusā un stilu līdzsvarotība: valodas un tulkojuma variantus var izpētīt tikai dažāda stila tekstos;

(3) dažādas sarežģītības pakāpes tekstu iekļaušana korpusā: tekstiem, kas ir iekļauti korpusā, jārāda reālā tulkošanas sarežģītība;

(4) nepārveidotu tekstu izvēle: tulkojuma fenomenu var pētīt tikai saistībā ar autentiskiem tekstiem;

(5) svarīgi, lai būtu pieejami bibliogrāfijas dati: lietotājiem jābūt pārliecinātiem, ka korpuss, kuru viņi izmanto, atbilst viņus interesējošai problēmai un ka tas nodrošina uzticamus un reprezentatīvus datus [Ulrych 1997:429–430].

Visvairāk pētījumu ir veikts, izmantojot angļu-norvēģu valodas paralēlo korpusu. Uz tā pamata ir skatītas gan tulkojuma tematiskās struktūras [Hasselgård 1998], gan tulkošanas procesa fāzes, gan galvenie tulkošanas atbilsmju tipi [Thunes 1998]. Tas viss rada fonu turpmākai paralēlo tekstu izmantošanai, kur vēl ir daudz nezināmu lietojumu gadījumu, jo, kā uzsver M. Beikere, „tulkošanas studijās korpusa metodes un korpusa būtības izmantošana vēl aizvien ir bērnu autiņos" [Baker 1999:281].

Terminoloģijas pētījumi. Līdz ar pieaugošo daudzvalodu dokumentu klāstu ikdienā ir nepieciešams veidot tulkošanas atmiņas un pētīt terminoloģiju. Paredzēts, ka divvalodu korpusā būs iekļauta dažādu funkcionālo stilu leksika, kas ļaus to izmantot terminoloģijas pētījumiem un terminoloģijas automātiskās izguves programmas apmācīšanai. Taču var veidot arī speciālu vienas jomas korpusu (vai, piemēram, juridisko tekstu datu bāzi[66]), kur šī speciālā terminoloģija būs pārstāvēta vēl pilnīgāk.

Pasaules prakse. Šodien paralēlie korpusi tiek plaši izmantoti, lai tulkotāji apgūtu attiecīgās nozares terminoloģiju. Zviedrijā, lai apmācītu tulkotājus, tika izveidots apakškorpuss, kura sastāvā bija tehniskās rokasgrāmatas, politiskie dokumenti un finanšu atskaites zviedru un angļu valodā (un vēl dažās citās valodās). Danielsone un Ridingss [Danielsson, Ridings 2000] min, ka šis apakškorpuss izraisīja daudz diskusiju par termina jēdzienu vispār. Pat tajos gadījumos, kad ir zināms termins avotvalodas tekstā, var būt zināmas grūtības atrast atbilstošo terminu mērķavota tekstā. Tas vēlreiz apliecina, ka paralēlie korpusi atspoguļo zināšanas par attiecīgo kultūru un ar nozari saistīto informāciju.

Pīrsone informē par terminoloģijas studijām Dublinas pilsētas universitātē, lai iedrošinātu studentus „izmantot drīzāk autentiskus tekstus, nevis speciālās vārdnīcas, lai iepazītu terminoloģiju un atpazītu terminu nozīmi un lietojumu" [Pearson 2000:92]. Galvenā autores uzmanība ir pievērsta terminoloģijas avotu kritērijiem un tam, kā notiek terminoloģijas izguve no tekstiem.

Valodu apguve un mācīšana. Gan vienvalodas, gan paralēlo korpusu var izmantot valodas apguvē un mācīšanā: veidojot gramatikas vingrinājumus, pilnveidojot leksikas apguvi, izstrādājot mācību grāmatas un vārdnīcas.

Pasaules prakse. Viens no nesenajiem ieguvumiem paralēlo un vienvalodas korpusu izmantošanai valodu apmācībā ir veiksmīgs teorētisko zināšanu apvienojums ar valodas praktisko pieredzi. Tas tiek panākts ar valodas korpusu palīdzību – skolēni (studenti) iegūst jaunus datus un iepazīst reālu valodas lietojumu. Lai to panāktu, ir nepieciešams:

[66] http://europa.eu.int/eur-lex/lex/lv/index.htm – skatīts 30.07.2005.

(1) augstas kvalitātes korpuss, kas ir pietiekami reprezentatīvs un pietiekami liels, lai nodrošinātu pienācīgu datu apjomu;

(2) implementētas, viegli lietojamas, elastīgas piekļuves, vaicājuma un analīzes procedūras [Peters et al. 2000:74].

Pēc Itālijas pētnieku domām, paralēlajiem un salīdzināmajiem korpusiem ir dažādi lietojumi valodas apguves un mācīšanas procesos. Paralēlos korpusus „veido reālās pasaules teksti, no kuriem var iegūt datus par kontekstuāliem tulkojuma ekvivalentiem" (ibid), tas noder parastam valodas studentam. Salīdzināmos korpusus „var apstrādāt, lai iegūtu informāciju par divu valodu leksisko ekvivalentu kontekstiem noteiktā jomā" (ibid), un tas der pieredzējušiem pētniekiem. Tādējādi Pizā ir izveidotas divas dažādas sistēmas – gan paralēlā korpusa sistēma, gan salīdzināmā korpusa sistēma.

Arī Vācijā, izmantojot angļu-vācu tulkojumu korpusu, ir iecere tā datus iestrādāt Hemnicas interneta gramatikā[67], kuras mērķauditorija būs gan valodnieki un pētnieki, gan studenti un pasniedzēji.

Dabīgās valodas apstrāde un mašīntulkošana. Tāpat kā vienvalodas korpusu, arī daudzvalodu korpusus plaši izmanto dabīgās valodas apstrādē – gan tekstu kopsavilkuma veidošanā, gan dažādu programmatūru izvērtēšanā, gan programmrīku izstrādē. Paralēlos korpusus izmanto arī daudzvalodu terminoloģijas izgūšanā, tādu sistēmu izstrādei izmanto sastatīšanu vārda līmenī, un šī tehnoloģija ir sevi pierādījusi ne tikai radniecīgu valodu tekstu apstrādē, bet arī, piem., angļu-japāņu un angļu-ķīniešu valodu analīzē.

Tā kā atšķirīgās valodās leksiskā un gramatiskā neviennozīmība atspoguļojas dažādā mērā, tad paralēlā korpusa dati var tikt analizēti neviennozīmības problēmu risināšanā. Veicot starpvalodu salīdzinājumu, attiecīgais konteksts var palīdzēt noteikt neviennozīmīgo vienību.

Mašīntulkošanas sistēmu izstrādē gan tieši, gan netieši ir nepieciešama informācija no divvalodu korpusiem. Vērtīgi būtu, ja teksts, kas sastopams divvalodu korpusā un ir sastatīts teikuma līmenī, būtu pieejams arī vienvalodas korpusā ar vārdšķiru un sintaktisko marķējumu. Tā iegūtu tekstu ar lielu pievienoto vērtību, kas ļautu uzlabot mašīntulkošanas sistēmu efektivitāti un palīdzētu risināt daudznozīmības problēmas.

Paralēlā korpusa datus var plaši izmantot daļēji automatizētas tulkošanas sistēmas izstrādē, piemēros balstītās un statistiskās mašīntulkošanas sistēmās, to apmācīšanā (piem., tulkošanas ekvivalentu novērtēšanā).

[67] http://www.tu-chemnitz.de/phil/InternetGrammar/publications/info/grammar.htm – skatīts 7.07.2005.

Pasaules prakse. Pirmie divvalodu korpusi pasaulē tika izveidoti tieši mašīntulkošanas vajadzībām (piem., Kanādas parlamenta dokumentu korpuss angļu un franču valodā *Canadian Hansard*[68]). Statistiskās mašīntulkošanas sistēmas vajadzībām tika izveidots teikuma līmenī sastatīts EUROPARL korpuss[69], tajā iekļauti 11 Eiropas valodu dokumenti no Eiropas Parlamenta. Tā apjoms ir aptuveni 28 milj. vārdlietojumu katrā valodā.

Sastatītu franču-angļu valodas korpusu BAF[70] izmantoja automatizētas tulkošanas atbalstam (tulkošanas atmiņas izveidošanai), kā arī juridisko tekstu apstrādē (kopsavilkuma veidošanai).

4.3. Runas korpusa izmantošana

Vienkopus uzkrājot un aprakstot plašu runāto latviešu valodu, tiek sagatavots materiāls gan teorētiskiem pētījumiem, gan praktiskām izstrādēm.

Runas korpuss ir nepieciešams saistītas runas izpētē un analīzē, un pētījumu rezultāti var tikt praktiski izmantoti latviešu valodas runas sintezatora un runas atpazīšanas sistēmu izstrādē, piemēram, mobilo un stacionāro telefonu tīklu uzziņu sistēmas, viesnīcu uzziņu un pasūtījumu sistēmas, teksta automātiska atskaņošana u. c. Pētījumi, kas balstīti uz plašu runas materiālu analīzi, ir nepieciešami, izstrādājot palīgierīces neredzīgiem vai vājredzīgiem cilvēkiem, piemēram, teksta-runas sintēzes sistēma, kas pārveido elektroniskā formā esošu tekstu (e-pasta vēstules, dokumentus, ziņu portāla rakstus u. tml.) skaniskā formā.

Runas korpuss būs noderīgs, veidojot latviešu valodas mācīblīdzekļus (gan datorizētus, gan drukātus) un latviešu valodas vārdnīcas, kurās norādīta latviešu valodas vārda izruna.

Veidojot latviešu valodas runas korpusu, tiks radīta mūsdienīga bāze tālākiem latviešu fonētikas (īpaši latviešu valodas prosodijas), sintakses pētījumiem. Runas korpuss ļaus meklēt atbildes uz līdz šim maz pētītiem jautājumiem:

(1) kādas fonētiskās pārmaiņas notiek vārdu sadurā (piem., patskaņu redukcija vārda beigās, asimilācija balsīguma ziņā);

(2) kā mainās vārda uzsvara un palīguzsvara vieta saistītā runā;

(3) vai iespējams izstrādāt algoritmu automatizētai zilbju intonāciju noteikšanai vārdos;

(4) kā automātiski ģenerēt prosodiju (piemēram, vārda uzsvaru, teikuma intonāciju) runas sintēzes sistēmās u. tml.

[68] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20 – skatīts 7.07.2005.
[69] http://people.csail.mit.edu/koehn/publications/europarl – skatīts 7.07.2005.
[70] http://rali.iro.umontreal.ca – skatīts 7.07.2005.

Parasti runas korpusa daļa tekstu korpusā veido nelielu daļu (pēc pasaules pieredzes apm. 10–15%), bet arī tas ir ļoti nozīmīgi valodas pētniecībai un citiem mērķiem, jo iespējams veikt pētījumus par intonāciju (prievārda konstrukciju, teikuma intonāciju); iespējams izvērtēt dažādas pieejas prosodijas anotēšanā, var skatīt teikuma intonācijas dažāda stila tekstos (piem., dievkalpojums un dzejas priekšlasījumi).

Pasaules prakse. Viens no pirmajiem runas korpusiem ir Londonas-Lundas britu angļu valodas korpuss[71], kas tapis apmēram tajā pašā laikā (izveidots 1975. gadā), kad visiem labi zināmie teksta korpusi: Brauna korpuss (*Brown corpus of American English*[72]) un Lančestras-Oslo/Bergenas korpuss (*Lancaster-Oslo/Bergen corpus of written British English*[73]). Londonas-Lundas korpusa veidotāju mērķis bija padarīt runātos tekstus pieejamus mašīnlasāmā formā: tika savākti un transkribēti 87 teksti (pavisam 435 000 vārdlietojumu). Korpusā ietvertie teksti ir transkribēti, bet ne gramatiski marķēti.

Pasaulē līdzās daudzmiljonu rakstītu tekstu korpusiem ir atrodami arī nedaudzi runātās (spontānās runas) valodas korpusi. Tiem ir milzīga nozīme runas tehnoloģiju izstrādē. Ir vairāki angļu valodas runas korpusi (*British National Corpus*[74] , *Santa Barbara Corpus of American English*[75], *CANCODE* korpuss[76] u. c.) un daži citu valodu runas korpusi, piem., nīderlandiešu (*Corpus Gesproken Nederlands*[77]), basku valodas runas korpuss[78]) u. c. Runas transkribēšana un marķēšana ir dārgs process. Pēdējās desmitgadēs ir izstrādāti daudzi marķēšanas un apstrādes rīki rakstītu tekstu korpusiem, bet nav pietiekami runas datu apstrādes rīki. Esošie programmrīki ne vienmēr ir izmantojami spontānas runas apstrādei un analīzei.

4.4. Valodas korpusa pieejamība

Valodas korpusi mēdz būt komerciālie vs. nekomerciālie valodas korpusi pētniecības mērķiem; tiešsaistes korpusi vs. korpusi, kurus izplata CD formātā.

Jebkura valodas korpusa lietotāji parasti aptver vairākas grupas (piem., lietotāji, kas korpusa datus izmanto tikai mācību un pētniecības mērķiem; lietotāji, kas datus apstrādā tālākai iekļaušanai komercprodukta izstrādē). Katrai lietotāju grupai ir dažādas pieejamības tiesības.

[71] http://khnt.hit.uib.no/icame/manuals/LONDLUND/index.htm – skatīts 21.07.2005.
[72] http://helmer.aksis.uib.no/icame/brown/bcm.html – skatīts 21.07.2005.
[73] http://khnt.hit.uib.no/icame/manuals/lob/index.htm – skatīts 21.07.2005.
[74] http://www.natcorp.ox.ac.uk – skatīts 21.07.2005.
[75] http://www.ldc.upenn.edu/Projects/SBCSAE – skatīts 22.07.2005.
[76] http://www.cambridge.org/elt/corpus/cancode.htm – skatīts 22.07.2005.
[77] http://www.elis.rug.ac.be/cgn/index_nl.html – skatīts 22.07.2005.
[78] http://www.elda.org/catalogue/en/speech/S0123.html – skatīts 22.07.2005.

Ja valodas korpuss paredzēts komerciāliem mērķiem, jebkura tā izplatīšana ir tikai par maksu.

Valodas korpuss pētniecības mērķiem var būt brīvi pieejams, taču var būt gadījumi, kad kāda korpusa daļa (piem., marķēts korpuss vai drīzāk – pilna korpusa versija), kā arī kādi speciāli lietojumi var būt pieejami par abonentmaksu. Abonentmaksa parasti ir atšķirīga privātpersonām un juridiskām personām. Kā liecina latviešu valodas korpusa iespējamo lietotāju aptaujas anketa, vairāk nekā trešdaļa respondentu (38%) nav gatavi maksāt abonēšanas maksu par elektroniskajiem latviešu valodas resursiem, bet gandrīz tikpat daudz respondentiem (35%) par to nav galīga viedokļa.

Pasaules prakse valodas korpusa pieejas nodrošināšanā ir dažāda: sākot ar lieliem valodas korpusiem, kas ir publiski pieejami bez maksas (piem., Lietuviešu valodas korpuss), beidzot ar brīvi pieejamām bezmaksas korpusu daļām. Ir korpusi, piem., Britu nacionālais korpuss, kuru arī pētniecības mērķiem var izmantot tikai par maksu.

Lielākā daļa korpusu ir izmantojami tiešsaistē, bet daļa korpusu tiek izplatīti arī ar kompaktdisku, piemēram, METER korpuss[79] vai Britu nacionālais korpuss.

Aptaujājot interesentus par ērtāko veidu, kā piekļūt elektroniskajiem resursiem, 37% respondentu atzina, ka vēlētos izmantot elektroniskos resursus tiešsaistē, 6% – ar kompaktdisku, 20% – lejupielādējot, 3% – tiešsaistē un ar kompaktdisku, 17% – tiešsaistē un lejupielādējot, 11% – tiešsaistē, ar kompaktdisku un lejupielādējot, bet 6% – ar kompaktdisku un lejupielādējot. Tas nozīmē, ka latviešu valodas korpusa veidotājiem ir jāparedz gan autortiesību, gan tehnoloģisko iespēju risinājums šādu iespēju nodrošināšanai. Sīkāk par autortiesību risinājumiem skatīt 6. nodaļu. Savukārt valodas korpusa izmantošanas iespēju tehnoloģiskos piedāvājumus skatīt 5. nodaļā.

Koncepcijas autori iesaka izveidot vispārīgu latviešu valodas korpusu pētniecības mērķiem, kas būs pieejams tiešsaistē lietotājiem, noslēdzot līgumu par izmantošanu mācību un pētniecības mērķiem.

Vēres

Altenberg B. [1998], "Connectors and sentence openings in English and Swedish." – *Corpora and cross-linguistic research. Theory, method, and case studies,* ed. by S.Johansson, S.Oksefjell. – Amsterdam: Rodopi – pp. 115–143.

Baker M. [1995], "Corpora in translation studies: An overview and some suggestions for future research." – *Target* 7(2). – pp. 223–243.

Baker M. [1999], "The role of corpora in investigating the linguistics behaviour of professional translators." – *International Journal of Corpus Linguistics,* Vol. 4(2). – pp. 281-298.

Biber D., Finegan E., Atkinson D. [1994], "ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers." – *Creating and*

[79] http://www.dcs.shef.ac.uk/nlp/meter/Metercorpus/metercorpus.htm – skatīts 28.07.2005.

Using English Language Corpora, ed. by U. Fries, G. Tottie, P. Schneider. – Amsterdam: Rodopi – pp. 1–14.

Biber D., Conrad S., Reppen R. [1998], *Corpus linguistics: investigating language structure and use.* – Cambridge: Cambridge University Press.

Danielsson P., Ridings D. [2000], "Corpus and terminology: software for the translation program at Göteborgs Universitet or getting students to do the work." – *Multilingual Corpora in Teaching and Research*. – Amsterdam: Rodopi – pp. 65–72.

Dickens A., Salkie R. [1996], "Comparing bilingual dictionaries with a parallel corpus." – *EURALEX'96 Proceedings.* – Göteborg University – pp. 551–559.

Ebeling J. [1998], "Using translations to explore construction meaning in English and Norwegian." – *Corpora and cross-linguistic research. Theory, method, and case studies,* ed. by S. Johansson, S. Oksefjell. – Amsterdam: Rodopi – pp. 169–195.

Fries U., Schneider P. [2000], "ZEN: Preparing the Zurich English Newspaper Corpus. " – *English Media Texts: Past and Present,* ed. by F. Ungerer. – Amsterdam: John Benjamins – pp. 1–24.

Gale W. A., Church K. W. [1993], "A Program for Aligning Sentences in Bilingual Corpora." – *Computational Linguistics,* Volume 19, Number 1. – pp. 75–90.

Gellerstam M. [1986], "Translationese in Swedish novels translated from English." – *Translation studies in Scandinavia,* ed. by L.Wollin, H.Lindquist. – Lund: CWK Gleerup – pp. 88–96.

Hasselgård H. [1998], "Thematic structure in translation between English and Norwegian." – *Corpora and cross-linguistic research. Theory, method, and case studies,* ed. by S. Johansson, S. Oksefjell. – Amsterdam: Rodopi – pp. 145–167.

Hofland K., Johansson S. [1998], "The translation corpus aligner: a program for automatic alignment of parallel texts." – *Corpora and cross-linguistic research. Theory, method, and case studies,* ed. by S. Johansson, S. Oksefjell. – Amsterdam: Rodopi – pp. 87–100.

Hofland K., Johansson S. [1982], *Word frequencies in British and American English.*– Bergen: Norwegian Computing Centre for the Humanities.

Johansson M. [1996], "Fronting in English and Swedish: A text-based analysis." – *Synchronic corpus linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*. – Amsterdam: Rodopi – pp. 29–39.

Johansson S., Hofland K. [1994], "Towards and English-Norwegian parallel corpus." – *Creating and using English language corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora*. – Amsterdam: Rodopi – pp. 25–37.

King P. [1997], "Parallel corpora for translator training." – *Practical Applications in Language Corpora,* ed. by B. Lewandowska-Tomaszcyk, P. James Melia. – Łódź: Łódź University Press – pp. 393–402.

Kortmann B., Herrmann T., Pietsch L., Wagner S., [2005], *A Comparative Grammar of English Dialects: Agreement, Gender, Relative Clauses*. – Berlin/New York: Mouton de Gruyter.

Pearson J. [2000], "Teaching terminology using electronic resources." – *Multilingual Corpora in Teaching and Research*. – Amsterdam: Rodopi – pp. 92–105.

Peters C., Picchi E., Biagini L. [2000], "Parallel and comparable bilingual corpora in language teaching and learning." – *Multilingual Corpora in Teaching and Research*. – Amsterdam: Rodopi – pp. 73–85.

Quirk R., Greenbaum S., Leech G., Svartvik J. [1985], *A Comprehensive Grammar of the English Language*. – London.

Rissanen M. [1992], “The diachronic corpus as the window to the history of English.” – *Directions in Corpus Linguistics*. – Berlin & New York: Mouton de Gruyter – pp. 185–205.

Schmied J., Schäffer H. [1996], “Approaching translationese through parallel and translation corpora.” – *Synchronic corpus linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*. – Amsterdam: Rodopi – pp. 41–56.

Schmied J., Fink B. [2000], “Corpus-based contrastive lexicology: the case of English *with* and its German translation equivalents.” – *Multilingual Corpora in Teaching and Research*. – Amsterdam: Rodopi – pp. 157–176.

Schmied J. [1998], “Differences and similarities of close cognates: English *with* and German *mit.*” – *Theory, method, and case studies,* ed. by S. Johansson, S. Oksefjell. – Amsterdam: Rodopi – pp. 255–275.

Thunes M. [1998], “Classifying translational correspondences.” – *Corpora and cross-linguistic research. Theory, method, and case studies,* ed. by S. Johansson, S. Oksefjell. – Amsterdam: Rodopi – pp. 25-50.

Ulrych M. [1997], “The impact of multilingual parallel concordancing on translation.” – *Practical Applications in Language Corpora,* ed. by B. Lewandowska-Tomaszczyk, P. James Melia. – Łódź: Łódź University Press – pp. 421–435.

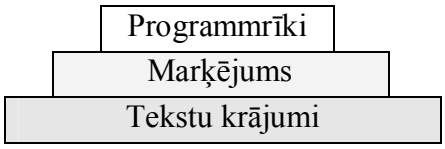
Viberg Å. [1998], “Contrast in polysemy and differentiation: Running and putting in English and Swedish.” – *Corpora and cross-linguistic research. Theory, method, and case studies,* ed. by S. Johansson, S. Oksefjell. – Amsterdam: Rodopi – pp. 343–376.

Wikberg K. [1996], “Evidence form the English-Norwegian Parallel Corpus.” – *Synchronic corpus linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*. – Amsterdam: Rodopi – pp. 17–28.

5. Latviešu valodas korpusa programmatūras izveides principu piedāvājums

Datu bāzē iekļaujamo tekstu raksturojums stilistiskā, hronoloģiskā u. c. aspektos; tekstu ievades principu noteikšana; mutvārdu un rakstveida tekstu apjoma noteikšana u. c. jautājumu risinājumi.

Būtiskāko pievienoto vērtību elektronisko tekstu krājumiem sniedz marķējums (metainformācija), kas padara tekstus ne tikai mašīnlasāmus (*machine-readable*), bet atkarībā no marķējuma līmeņa arī mašīnai saprotamus (*machine-understandable*). Marķējums ir formāla (iezīmju) valoda, ar kuras palīdzību tiek veidots starpslānis starp tekstiem brīvā formā un programmrīkiem – korpusa lietojumiem (sk. 5. 1. attēlu). Balstoties uz papildinošo metainformāciju, kļūst iespējams veidot lietojumus, kas nodrošina tekstu struktūras un satura automatizētas analīzes iespējas dažādos aspektos.



5.1. attēls – marķējums ir formāls starpslānis starp tekstu saturu (valodu) un korpusa funkcionālajiem lietojumiem.

5.1. Izstrādes vadlīnijas

Koncepcijas autori iesaka ņemt vērā šādas vadlīnijas:

1) tekstu aprakstīšanas, strukturēšanas un marķēšanas standartizācija; vispārpieņemtu standartu izvērtēšana un izmantošana (vajadzības gadījumā – kombinēšana un paplašināšana);
2) sistēmai (tekstu bāzei un programmrīkiem) jābūt viegli pieejamai/izplatāmai ar pēc iespējas minimālām lietotājdatora un programmatūras konfigurācijas prasībām;
3) tā kā nākotnē radīsies nepieciešamība pēc arvien jauniem korpusa funkcionālajiem lietojumiem, vienlaikus arī programmrīkiem, kā arī var mainīties realizētās funkcionalitātes prasības, korpusa lietojums un datu modelis jāveido vienots, viegli uzturams un paplašināms.

5.2. Datu modelis

Datu modeļa projektēšana ir ļoti svarīgs izstrādes posms. Tas ir pamats, uz kura balstās visa korpusa lietojumu izstrāde, un no tā ir atkarīga veiksmīga kopējās korpusa sistēmas realizācija, kā arī vēlāko izmaiņu, uzlabojumu un paplašinājumu ieviešanas sarežģītība un elastība. Korpusa gadījumā par datu modeli bez datubāzes shēmas ir jāuzskata arī dažādo marķējuma līmeņu gramatikas. Lai nodrošinātu korpusa savietojamību ar dažādiem atšķirīgu autoru kolektīvu izstrādātiem programmrīkiem un

atvieglotu gramatiku definēšanu un turpmāku to paplašināšanu, par pamatu ir jāņem vispārpieņemti (tekstu korpusu) marķēšanas standarti (atkarībā no līmeņa).

Datorlingvistikā ir izšķirami vairāki tekstu marķēšanas līmeņi:

0. **līmenis**: nemarķēts teksts – tāds teksts, kuram ir tikai dabīgās iezīmes (pieturzīmes, rindu pārnesumi, lielie burti, saīsinājumi u. c.). Tas ir vienkāršs teksts bez pievienotās vērtības – jebkāda veida datorlingvistiskās apstrādes un formālām anotācijām; šādu tekstu varētu iegūt, vienojoties ar dažādiem plašu elektronisko resursu turētājiem;

1. **līmenis**: strukturālā marķēšana – šajā līmenī tekstam tiek aprakstīta satura loģiskā struktūra, piemēram, tiek norādītas virsrakstu, nodaļu, rindkopu, teikumu, tiešo runu, teicēju, citātu svešvalodu fragmentu u. c. loģiskās struktūras elementu robežas.

Īpašs šī gadījuma apakšlīmenis ir reprezentatīvas, procedurālas iezīmes, piemēram, trekināts teksts slīpraksts, atkāpes u. tml. (HTML, PDF u. c. formāti) tekstu publicēšanas nolūkiem; šādi nav tieši aprakstīta tekstu uzbūve, bet tikai to vēlamais vizuālais noformējums.

2. **līmenis**: morfoloģiskā marķēšana – attiecīgo vārdu raksturojošās morfoloģiskās informācijas formāla pievienošana katram vārdam tekstā;

3. **līmenis**: sintaktiskā marķēšana – teikuma struktūras un locekļu formāla aprakstīšana;

4. **līmenis**: semantiskā marķēšana – vārdu semantiskais vai leksiski semantiskais raksturojums, piemēram, norādot īpašvārdus, verbu kauzativitāti. Izmantojot leksisku latviešu valodas taksonomiju, var sniegt sīkāku vārdu raksturojumu, piemēram, personas apzīmējums, etnonīms, dzīvnieks, augs; norādīt hiperonīmiskās un sinonīmiskās u. c. attieksmes.

Runas korpusa specifiskā marķēšana, kas lietojama papildus rakstītā teksta marķēšanai:

1) tekstu transkribēšana, ortogrāfiskā atveide un prosodijas marķēšana;
2) fonētiskā transkripcija.

Atsevišķi izdalāms metadatu veids pretstatā satura elementu (tekstu iekšējās uzbūves) anotēšanai ir ārējā metainformācija, t. i., netieša, bet saistīta, aprakstoša informācija, vispārīgi dati, kas attiecas uz katru konkrēto tekstu kopumā: autors, žanrs, publicēšanas datums, autortiesības, teksta (elektroniskās versijas) statuss, ticamība

(kvalitāte) utt. Tekstu korpusā tas nodrošina paplašinātas meklēšanas iespējas un tekstu klasificēšanu/ierobežošanu dažādām apstrādes vajadzībām. Šajā koncepcijā ar metadatiem tiek saprasti tekstu ārējie metadati, bet ar marķējumu – iekšējie.

Papildus skatīt 5. 2. attēlu.

5.2.1. *Metadati*

Metadatu struktūra neatkarīgi no valodas korpusa apakškopas un teksta veida ir nepieciešama viena un tā pati, tādēļ svarīga ir vienota standarta izveidošana, par pamatu ņemot vispārēju standartu. Priekšrocība: savietojamība un metadatu saprotamība starp dažādiem jau esošiem programmrīkiem, interneta meklēšanas servisiem u. tml. Daži populārākie (tekstu) metadatu aprakstīšanas standarti ir īsi aprakstīti turpmāk tekstā. Latviešu valodas korpusa vajadzībām metadatu gramatikā var tikt apvienoti elementi no dažādiem standartiem, tā var tikt paplašināta ar specifiskiem elementiem, taču ir jāsaglabā automātiskas transformēšanas iespējas uz dažādu standartu modeļiem.

Vieni no izplatītākajiem starptautiskajiem standartiem ir šādi:

- **TEI** (*Text Encoding Initiative*) ***Header***[80]**;**
- **CES** (*Corpus Encoding Standard*) ***Header***[81] – *TEI Header* paplašinājums, piemērojot to tekstu korpusu speciālām vajadzībām;
- **DC**[82] – *Dublin Core Metadata Element Set*. DC pamatā ir 15 elementu kopa, kuru var iedalīt trijās kategorijās, aprakstot saturu, intelektuālo īpašumu un elektroniskā formāta versijas. Savietojams ar TEI *Header*, taču veidots vienkāršāks un vispārīgāks. DC ir orientēts tīmekļa resursu aprakstīšanai, tam ir specificēta RDF/XML shēma semantiskā tīmekļa meklēšanas robotu un aģentu vajadzībām;
- **EAGLES/ISLE** ***Meta Data Initiative***[83] – multimediju/multimodālu valodas resursu metadatu aprakstīšanas standarts; balstīts uz TEI, CES u. c. esošiem standartiem, paplašinot tos.

Dažādu kategoriju informācija, ko metadatiem vajadzētu aptvert:

1) informācija, kas tiešā mērā attiecas uz teksta vienību: darba nosaukums, autors, žanrs, tematika, valoda/dialekts/izloksne, tapšanas vieta (ja zināma), īsa anotācija, autora dzimums u. c.;

[80] http://www.tei-c.org/P4X/HD.html – skatīts 4.07.2005.
[81] http://www.cs.vassar.edu/CES/CES1-3.html – skatīts 4.07.2005.
[82] http://dublincore.org – skatīts 4.07.2005.
[83] http://www.mpi.nl/world/ISLE – skatīts 4.07.2005.

2) bibliogrāfiskie dati: izdevējs, krājums, informācija par tulkojumu, publikācijas veids, izdošanas vieta un gads, u. c.;
3) informācija par elektronisko formātu, pievienoto vērtību, autortiesībām: vārdlietojumu skaits, iestāde/persona, kas sagatavojusi tekstu; projekta vadītājs; marķējuma līmeņi; versija, datums; norāde, vai teksts ir manuāli pārbaudīts, vai tekstam piemērotais marķējuma līmenis ir manuāli pārbaudīts; autortiesības u. tml.;
4) runas korpusa gadījumā: runātāju skaits, vecums, dzimums, dzīves vieta u. c.
5) metadatu apakšgadījums ir katras teksta vienības piekļuves un operāciju tiesību raksturojums. Piemēram, teksts ir brīvi izmantojams vai ar ierobežotu pieejamību, kādas operācijas un kādi datu apjomi ir atļauti katrai no lietotāju kategorijām u. tml. Šie dati būs nepieciešami korpusa gala lietojumrīkiem, ņemot vērā lietotāja konta datus (kāds lietotājs veic darbības ar rīkiem).

Metadatu formalizēšanas piemērs ir atrodams 1. pielikumā.

5.2.2. *Marķējums*

Principā katrs no iepriekš minētajiem marķējumu līmeņiem kalpo kā ieejas dati nākošo līmeņu iegūšanā; nebūtisks izņēmums ir prezentācijas marķējums (0.5. līmenis), kuru principā var uzskatīt par 1. līmeņa apakšgadījumu (sk. 5. 2. attēlu). Pareizi veidotā korpusa sistēmā tas ir automātiski ģenerējams, izmantojot strukturālo (1. līmeņa) informāciju, prezentācijas līmeņa sagatavošanā nav jāiegulda nekāds manuāls darbs.

Strukturāli marķēti teksti (1. līmenis) ir iegūstami automatizācijas ceļā, analizējot 0. līmeņa tekstus, kā arī jau esošos manuāli sagatavotos 0.5. līmeņa tekstus (lielākoties HTML formā). Šim procesam LU MII jau ir iestrādes. Nepieciešama detalizēta metodika, izsmeļoša pieņēmumu, likumu un avotspecifisku šablonu kopa strukturālo elementu atpazīšanai un strukturālajai gramatikai atbilstošu rezultātu iegūšanai. Rezultāts ir manuāli jāpārbauda.

Morfoloģiskā marķējuma (2. līmenis) nodrošināšana un atbilstošas (daļēji) automatizētas marķēšanas sistēmas izstrāde ir viens no aktuālākajiem tuvākajā nākotnē izvirzāmajiem uzdevumiem. Latviešu valodas morfoloģiskā analizatora iestrādes jau ir, bet ir nepieciešama teorētisko (valodniecisko) nostādņu metodoloģijas pilnveidošana un algoritmu uzlabošana/paplašināšana. Iteratīva analizatora attīstīšana, sasniedzot

minimālu, apmierinošu analīzes kļūdu procentu. Galvenā problemātika ir saistīta ar daudznozīmības novēršanu.

Noteikta apjoma morfoloģiski marķētu, manuāli pārbaudītu tekstu krājuma izveide (apmēram 100 000 vārdlietojumu; šobrīd LU MII šādi ir marķēti ~15 000 vārdlietojumu), uz kā pamata varētu „apmācīt" automātisku morfoloģiskās marķēšanas rīku.

Sintaktiskā marķēšana. 3. marķēšanas līmenis, teikumu analīze un sintaktisko koku formalizēšana, praktiskā kvalitātē varētu būt pieskaitāms pie valodas korpusa izveides tālākajiem mērķiem, taču iestrāžu attīstīšana un eksperimentāli rezultāti ir tuvākās nākotnes uzdevumi.

Semantiskā marķēšana. 4. marķēšanas līmenis ir faktiski pēdējais un, domājams, ka būtu paredzams nākotnē. Abi pēdējie līmeņi ietver: metodikas un algoritmu izveidi, atbilstošu programmrīku projektēšanu un implementāciju, rezultātu manuālu pārbaudi/rediģēšanu.

Runas korpuss: tekstu fonētiskā **transkribēšana** un **prosodijas marķēšana**.

Runas transkribēšana ir diskursa atšifrēšana un fiksēšana mašīnlasāmā formā (parasti ortogrāfijā, retāk – fonētiskajā transkripcijā), norādot, piemēram, pauzes, ieelpas un izelpas vietas, runātāju maiņu sarunas laikā.

Transkripcijā ir svarīgi atspoguļot runas saturu, kvalitāti (piem., prosodiju, pauzes, balss kvalitāti), adresātu, kādi ekstralingvistiskie apstākļi ietekmē teikto (piem., apkārtne, darbība, sarunas dalībnieku raksturs un savstarpējās attiecības, attieksme vienam pret otru). Transkripcijai jāparāda temporālie aspekti (piem., pauzes ilgums, notikumu secība, vienlaicīga runāšana vai secīga runātāju maiņa), arī komentāri un paskaidrojumi. Vispārējā transkripcijā (*broad transcription*) aprakstītas vispārējas diskursa detaļas (līdzīgi kā scenārijos, lugās, tiesu sēžu protokolos), savukārt sīka transkripcija (*narrow transcription*) raksturo, piemēram, uzsvaru, intonāciju, žestus, balss kvalitāti (aizsmakusi, čerkstoša u. tml.), izrunas fonētiski fonoloģiskās īpatnības u. tml. Tas, cik vispārīga vai sīka būs runas transkripcija, ir atkarīgs no korpusa veidotāju mērķiem un uzdevumiem, protams, arī no finansējuma.

Transkribētu runu (tāpat kā teksta korpusu) ir iespējams strukturāli un morfosintaktiski marķēt, tāpēc ir jāizvēlas marķējums, kas būtu savietojams ar teksta korpusu.

Mācību korpuss: mācību korpusā uzkrātajos valodas studentu rakstu darbos tiek īpaši marķētas pieļautās kļūdas: pareizrakstības, morfoloģijas, sintakses, leksikas un stilistikas kļūdas, piem., nepareiza priedēkļverba vai sinonīma izvēle.

Standarti: TEI[84], CES[85], IPA[86] (*International Phonetic Alphabeth*) – starptautiskais fonētiskais alfabēts – un tā atveide mašīnlasāmā formātā – SAMPA[87] (*Speech Assessment Methods Phonetic Alphabet*) – u. c.

5.2.3. *Tehnoloģijas*

XML (*eXtensible Mark-up Language*) universāla marķēšanas valoda specializētu marķēšanas valodu definēšanai; vienkāršota SGML apakškopa (*Standard Generalized Markup Language*; vispārējs *de facto* standarts tekstuālas informācijas marķēšanā). XML formāts ir vienkārši lietojams un saprotams cilvēkam, vienlaicīgi ļoti elastīgi un efektīvi apstrādājams ar datorprogrammu palīdzību. Turklāt XML nav tikai marķēšanas valoda, bet arī vesels to atbalstošo tehnoloģiju un rīku kopums, kas nodrošina plašu funkcionalitāti:

- datu struktūru atbilstības pārbaudi attiecībā pret loģiskajiem uzbūves likumiem (DTD gramatikas, XML shēmas);
- navigāciju datu struktūrā – elementu adresāciju un atlasīšanu pēc norādītiem kritērijiem vai atrašanās vietas (*XPath* – XML *Path Language*);
- automātiskas datu transformācijas iespējas (XSLT – XML *Stylesheet Language for Transformations*) citos formātos, piemēram, HTML, PDF vai atbilstoši citai XML gramatikai;
- vēl daudz citu XML apstrādes iespēju, kurām jau ir izstrādātas vai top Tīmekļa konsorcija (W3C)[88] rekomendācijas (piem., *XPointer* elementu pozicionēšanai un *XQuery* vaicājumu valoda).

XML formāts ir īpaši piemērots kokveidīgu datu struktūru aprakstīšanai, un tekstu korpusa saturs (metadati + teksti, to struktūra) jau pilnīgi dabīgi veido šādu uzbūvi. XML ir platformneatkarīgs formāts, tas atbalsta *Unicode* rakstzīmju kodēšanas standartu un nodrošina ērtu datu apmaiņu/transformēšanu starp dažādām sistēmām.

DTD (*Document Type Definition*) ir formāla gramatika, ar kuras palīdzību tiek definēta XML dokumenta, šajā gadījumā – teksta – struktūra un tās primitīvo elementu datu tipi. Komplicētāka strukturēšanas gramatiku aprakstīšanas tehnoloģija ir XML

[84] http://www.tei-c.org – skatīts 05.07.2005.
[85] http://www.cs.vassar.edu/CES – skatīts 05.07.2005.
[86] http://www2.arts.gla.ac.uk/IPA – skatīts 05.07.2005.
[87] http://www.phon.ucl.ac.uk/home/sampa – skatīts 05.07.2005.
[88] http://www.w3.org – skatīts 05.07.2005.

Schema, kas dod iespēju veidot sarežģītākas datu struktūras un definēt detalizētākus elementu datu tipus. Datiem XML marķējumā ir jāatbilst vienai no divām kvalitātes pakāpēm:

1) strukturāli pareizi marķēti dati (*well-formed*), kas atbilst minimālajiem XML sintakses nosacījumiem – katrai atverošajai iezīmei ir atbilstoša aizverošā iezīme, un iezīmes ir savstarpēji pareizi ievietotas;
2) gramatiski pareizi (*valid*) strukturēti dati – iekļauj visus pirmās pakāpes nosacījumus un svarīgu papildu nosacījumu: iezīmes atbilst semantikai (loģiskajai uzbūvei), kas ir norādīta DTD gramatikā.

Attiecībā uz marķējuma līmeņiem, gramatika (shēma) ir izvēlētais/pielāgotais marķēšanas standarts (TEI, CES, DC u. c.), bet XML valoda – formāts jeb pieraksta veids. Piemērs strukturālai gramatikai un atbilstoši marķētam teksta fragmentam XML formātā ir apskatāms 2. pielikumā.

5.3. Programmrīki

Funkcionālam darbam ar korpusu ir nepieciešama atbilstošu programmrīku kopa. Korpusa programmatūru nekādā ziņā nevar uzskatīt par neatkarīgu, nesaistītu, atsevišķu rīku kopumu, bet gan par integrētu programmatūras sistēmu, kuras pamatā ir kopējs datu modelis. Tiesa, ņemot vērā vienotas vadlīnijas, korpusa komponenti var tikt izstrādāti neatkarīgi. Programmatūras komponentus var klasificēt, skatoties no dažādiem aspektiem (sk. turpmākās sadaļas). Vispārīgi runājot, korpusa sistēmai ir jānodrošina:

– tekstu glabāšana standartizētā, marķētā un apstrādāšanai ērtā formātā;
– navigācijas un meklēšanas iespējas tekstu krājumos,;
– specifiski lietojumi: valodnieciskā statistika, konkordances, vārdu savienojumu analīze u. c.;
– lietotāju saskarne darbam ar korpusu un tā lietojumiem.

5.3.1. *Centralizēta vs. decentralizēta korpusa sistēma*

Tiešsaistes (*on-line*) un bezsaistes (*off-line*) priekšrocības un trūkumi, kā arī iespējamās problēmas ir apskatītas koncepcijas 7. nodaļā.

5.3.2. *Lietotāju kategorijas*

Korpusa potenciālos lietotājus var iedalīt trīs galvenajās kategorijās:

– **akadēmiskie lietotāji** – valodnieki, literāti, translatoloģijas speciālisti, korpuslingvisti un citi, kas, lietojot specifiskus programmrīkus, tekstu

korpusu izmantos zinātniski pētnieciskiem, izglītojošiem un nekomerciāliem mērķiem;

- **komerciālie lietotāji** – pārsvarā institūcijas, kas korpusa saturu izmantos komerciāliem mērķiem, piemēram, izdevniecības vārdnīcu sastādīšanā un literāro darbu tulkošanā, tulkošanas aģentūras, IT firmas, kas nodarbojas ar valodas apstrādi, programmu testēšanā, apmācībā un lingvistisku datu bāzu izveidē (piem., pareizrakstības un sinonīmu vārdnīcas);
- **interesenti** – pārējie lietotāji, kas korpusam pieslēgsies fragmentāri, lai, visdrīzāk noskaidrotu kāda valodas lietojuma nozīmi, apkaimi.

Ārpus iepriekšminētajām atsevišķu kategoriju veido korpusa **administratīvie** lietotāji, kas nodarbosies ar tekstu apstrādi (piem., marķēšanu, indeksēšanu) un korpusa satura papildināšanu un rediģēšanu.

Attiecībā uz korpusa auditorijas sadalījumu lietotāju skaita ziņā paredzams, ka lielāko daļu aptvers akadēmisko lietotāju kategorija un interesenti, bet ievērojami mazāk lietotāju būs no komerciālā sektora. Tomēr ar laiku, papildinot korpusa saturu un palielinot tā pievienoto vērtību, kā arī attīstot programmrīku funkcionalitāti, šīs kategorijas nozīme varētu pieaugt.

5.3.3. *Korpusa satura izveide un vadība*

- Daļēji automatizēts, konfigurējams tekstu **strukturālās marķēšanas rīks**.
- Automātisks, konfigurējams **reprezentācijas** līmeņa **transformāciju rīks**.
- Latviešu valodas **morfoloģiskais analizators**.
  - Homonīmijas un daudznozīmības problēmu risināšana.
- Daļēji automatizēts **morfoloģiskās marķēšanas rīks.**
  - Lemmatizators: tiek iegūta vārda pamatforma (lietvārdiem – parasti vienskaitļa nominatīva forma (izņēmumi, piem., ģenitīveņi, daudzskaitlinieki u. tml.), darbības vārdiem – nenoteiksme).
- Daļēji automatizēts **sintaktiskās marķēšanas rīks**.
- Daļēji automatizēts **semantiskās marķēšanas rīks**.
- Tekstu **sastatīšanas** teikumu līmenī programmatūra (paralēlo korpusu izveidei). Precizitātes uzlabošana, izmantojot statistikas elementus u. c. līdzekļus.

– Tekstu krājumu **indeksēšana** (vārdformu/vārdlietojumu/biežuma/inversie u. c. indeksi).
– **Metadatu vadība**: pievienošana, papildināšana, rediģēšana.
– Paralēlajiem korpusiem – tekstu **sastatīšanas rīks**.
– **Atgriezeniskā saikne** ar lietotājiem – komentāri, pētījumu rezultāti, zinātnisko darbu publicēšana u. tml.; nepieciešama korpusa administrācijas kontrole; šis lietojums dotu arī potenciāli vērtīga satura papildināšanas iespēju un celtu tekstu pievienoto vērtību. Lietotāju uzvedības analīze (biežāk veikto vaicājumu tipi, biežāk lietotie statistikas dati, populārākās tekstu kategorijas/avoti u. tml.).
– Svarīgi ir nodrošināt ērtas un elastīgas korpusa **satura papildināšanas** un rediģēšanas iespējas, tekstu un marķējuma **versiju kontroli**. Lietotāju atgriezeniskās saiknes rezultātu uzturēšana.
– Centralizētai sistēmai: **pieejas kontroles sistēma**, autortiesību aizsardzības nodrošināšana. Brīvi pieejami teksti un teksti ar ierobežotām piekļuves/izmantošanas atļaujām. Lietotāju kontu/autorizācijas vadība. Lietotāju kategorijas. Piemēram, interesentiem ir pieeja tikai autortiesību neaizsargātiem tekstiem ar ierobežotu rīku funkcionalitāti, savukārt akadēmiskie un komerciālie lietotāji var strādāt gan ar publiskajiem tekstiem, gan ar individuālām pieejas tiesībām, arī ar tekstiem ar īpašu pievienoto vērtību; lietotāju autorizēta pieeja sniegtu korpusa uzturētājiem arī vērtīgu statistisku informāciju par lietotāju aktivitātēm, populārākajiem tekstiem un funkcionālajiem lietojumiem, lietotāju grupu īpašībām u. tml.

5.3.4. *Korpusa lietojumrīki*

– **Navigācijas sistēma**: hierarhiska korpusa (brīvi pieejamu) tekstu/fragmentu atlasīšana dažādos šķērsgriezumos, izmantojot **metadatus** un satura strukturālos elementus.
– Universāla, interaktīva, detalizēta **meklēšanas sistēma**; tekstu apgabala ierobežošana pēc metadatu kritērijiem un/vai dažādu līmeņu marķējuma elementiem; precīzi vai daļēji norādītu atslēgvārdu meklēšana, meklēšana, norādot šablonus, vārdu savienojumu un leksikosintaktisku šablonu norādīšana ar regulāro izteiksmju palīdzību u. c.
– **Konkordanču sistēma**, meklēšanas sistēmas paplašinājums; meklēšanas rezultātā atrastie vārdlietojumi tiek izcelti vidū, un tiem abās pusēs tiek parādīta konteksta apkaime (plaši tiek izmantots vārdnīcu veidošanā un tekstu tulkošanā); apkaimes izmērs – rakstzīmju skaits – lietotāja definējams vai ierobežots; kārtošana pēc labā/kreisā konteksta u. c.

Paralēlajiem korpusiem – paralēlās konkordances (potenciālo tulkošanas ekvivalentu meklēšana u. c. mērķi).

– **Kontekstu pozicionēšana** – meklēšanas un konkordances rīki lietotājam kā rezultātu dod norādes uz atrasto vārdlietojumu atrašanās vietu korpusā; pozicionēšanas rīks pēc lietotāja pieprasījuma apstrādā šādu norādi un nosūta lietotājam atbilstošo kontekstu; kontekstu izmērus ierobežo noteiktas strukturālo (konteineru) elementu robežas, kuras lietotājs var modificēt, ja to pieļauj konkrētās teksta vienības ierobežojumi; norādēm jābūt vienotā intuitīvi uztveramā formātā, lai tās var izveidot arī pats korpusa lietotājs.
– **Tekstu prezentēšana** – jebkura pieejamā teksta fragmenta (piem., pozicionēšanas konteksta) automātiska sagatavošana prezentācijas formātā (HTML, PDF); jāņem vērā pieejas un autortiesības. Arī pašiem vaicājumu rezultātiem (piem., konkordancēm) jābūt konvertējamiem dažādos formātos (PDF, XML u. c.) turpmākai izmantošanai arī ārpus korpusa sistēmas.
– **Sintakse**: sintaktisko koku vizualizācija, meklēšanas iespēju nodrošināšana pēc valences, leksikosintaktiskiem šabloniem.
– **Statistikas rīki**. Tos var iedalīt divās grupās:
    - gramatiskā un leksiskā statistika lietotāja izvēlētā teksta apgabalā, norādot arī citus kritērijus (piem., augšējās/apakšējās robežas, stopvārdu saraksti); sākotnēji – vārdformu un vārdlietojumu absolūtie/relatīvie biežumi u. c., nākotnē, attīstot tekstu augstāka līmeņa marķēšanu, nepieciešami sarežģītāki un pilnīgāki valodnieciskās statistikas rīki (piem., morfoloģiski marķētā korpusa daļā – lemmu līmenī); n-grammu analīze: vārdu/burtu savienojumu statistiskā (biežumu) analīze;
    - statistika par sistēmu, korpusa saturu un lietotājiem.
– Papildus ir jāparedz arī citi korpusa lietojumi: dinamiska, parametrizējama vārdu sarakstu, inverso vārdnīcu izguve u. c.
– Individuālām vēlmēm pielāgojama **kārtošana** un **grupēšana**: šis nav atsevišķs lietojums, bet svarīga funkcionālā īpašība praktiski ikvienam gala lietojumam: vārdu sarakstu kārtošana alfabēta augošā vai dilstošā secībā, grupēšana pēc stiliem, teritoriālā principa, hronoloģiski (piem., pa piecgadēm), statistiskās analīzes un meklēšanas rezultātu kārtošana pēc viena vai vairākiem kritērijiem (alfabētiski, pēc biežuma, pēc labās/kreisās

konkordances apkaimes, grupēšana, izmantojot metadatus u. tml.), navigācija dažādos šķērsgriezumos u. c.

– Pašreizējo valodneatkarīgo, brīvi pieejamo lietojumrīku, piemēram, Masarika Universitātē (Čehija) izstrādātā korpusu pārlūkošanas un vaicājumu rīka *Bonito*[89], izpēte, (marķēšanas) standartu atbalsts un izmantojamības iespējas latviešu valodas korpusa vajadzībām.

5.4. Tekstu ievades principi

Tekstu korpusa iespējamie uzkrāšanas veidi:

1) iegūstot esošus tekstus elektroniskā formā – panākot vienošanos (arī par autortiesībām) ar dažādām izdevniecībām, iestādēm, privātpersonām un citiem, kuriem ir valodas korpusam atbilstoši resursi; arī periodiski lejupielādējot atbilstošus tekstus no interesējošām tīmekļa vietnēm;
2) skenējot – teksti būs arī jāpārlasa un jāizlabo ievadkļūdas. Šāds process ir laikietilpīgāks par iepriekšminēto uzkrāšanas veidu, tomēr mūsdienās ir panākta gandrīz 98% tekstu ievadīšanas precizitāte. Domājams, ka šādi varētu tikt ievadīta daiļliteratūra u. c.;
3) manuāli ievadot – tādējādi valodas korpusam tiks pievienoti tādi teksti, kas nepieciešami reprezentativitātes nodrošināšanai un kuru formāts nav viegli (efektīvi) skenējams.

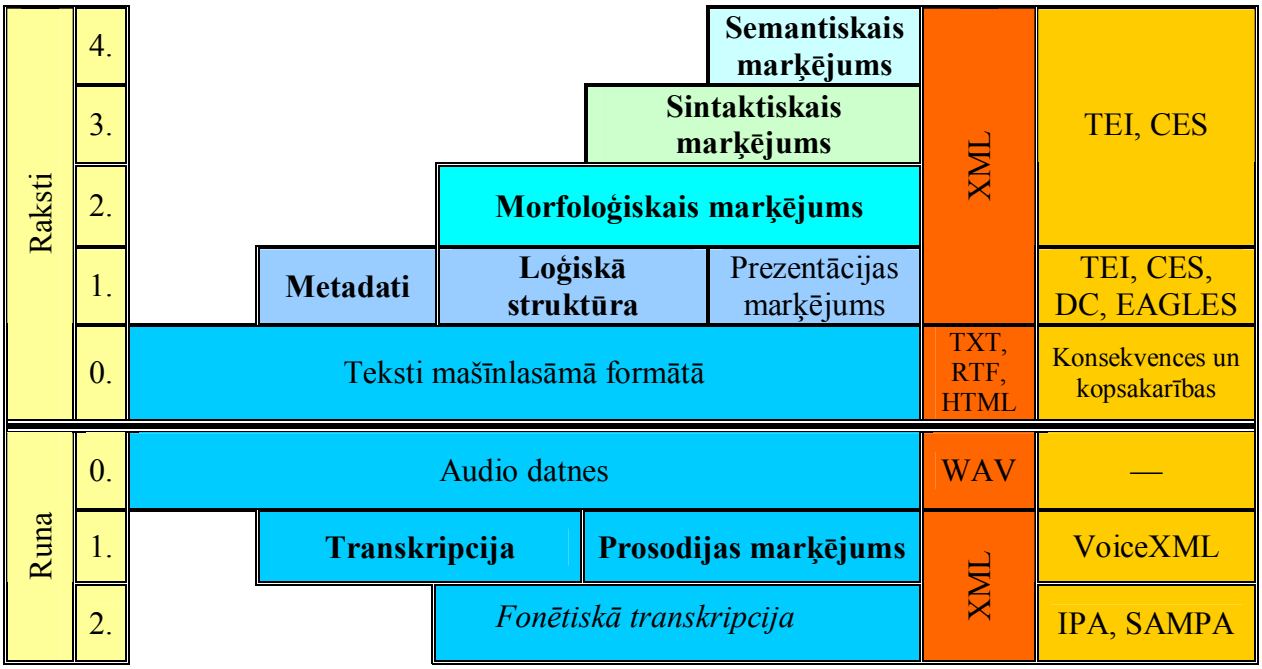
5.5. Kopsavilkums

Izstrādes darbi pa līmeņiem ir atkarīgi tikai vertikāli uz augšu, t. i., augstāku līmeņu ieguve ir atkarīga no zemāko līmeņu sagatavošanas, bet ne otrādi. Korpusa marķēšana un gala lietojumu izstrāde var norisināties (un tai būtu jānorit) vairākās iterācijās; aina, kas parādīta 5. 2. attēlā, visu laiku saglabājas. Iterācijas dažādu līmeņu ietvaros var notikt neatkarīgi. Arī programmatūras (gan satura vadība, gan gala lietojumi) attīstībai jānorit iteratīvi; korpusa lietojumi ir jāattīsta kopsolī ar marķējuma līmeņu attīstību (sk. 5. 1. tabulu). Sistēmas konceptuālais modelis vienkāršoti ir atainots 5. 1. diagrammā (UML – *Unified Modeling Language* – notācija).

Marķēšana:

1) manuāli (eksperimentāla marķēšana dažādu problēmu atklāšanai);
2) daļēji automatizēti (ir iestrādes un/vai produkti LU MII un sabiedrībā „Tilde”);
3) automātiski marķētāji, kas mācās no iepriekš samarķētajiem tekstiem.

[89] http://nlp.fi.muni.cz/projects/bonito – skatīts 21.07.2005.



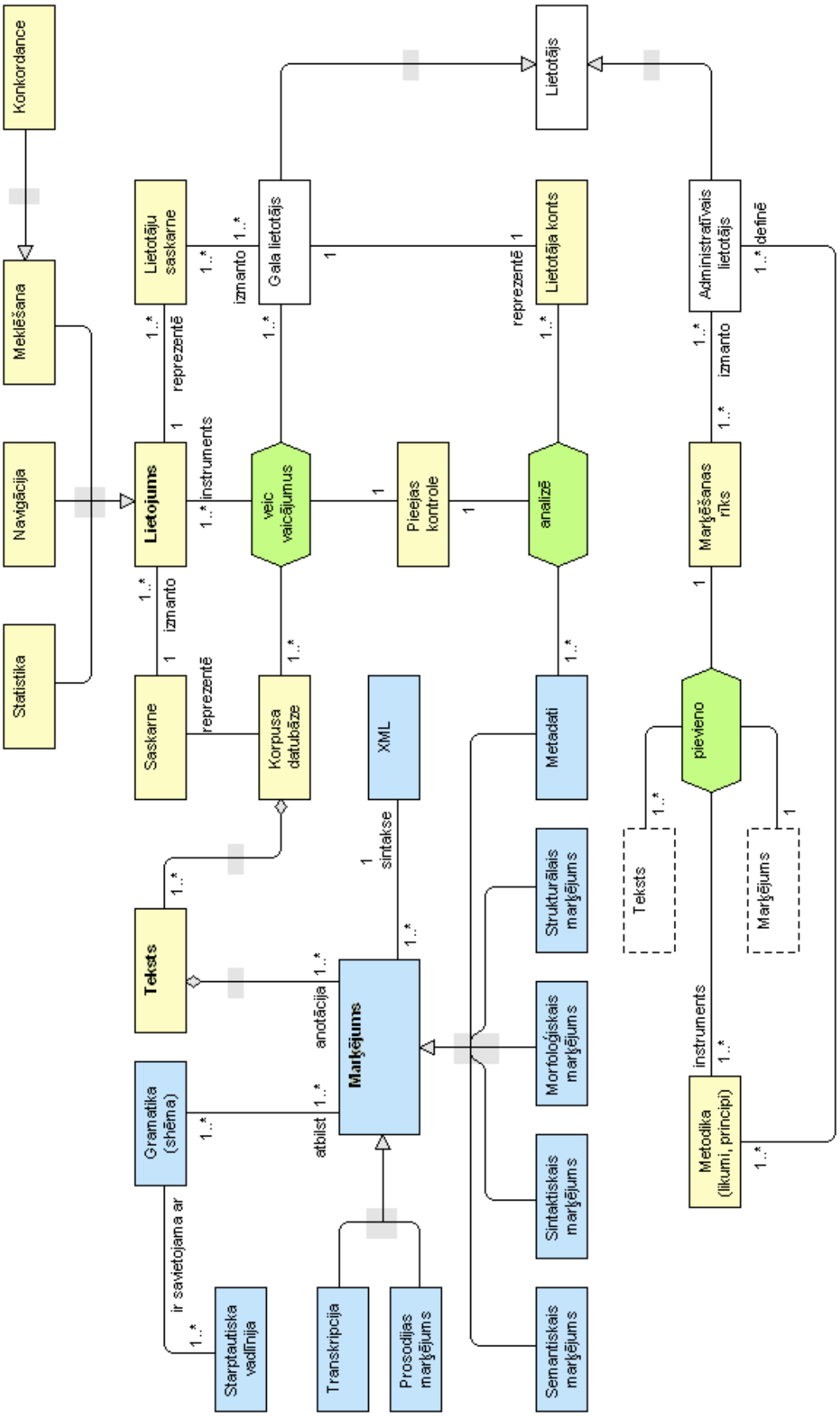
5.2. attēls – marķējumu līmeņu hierarhija un standarti.

Veicamās programmatūras izstrādes, sasniedzot noteiktu līmeni korpusa marķējumā, ir parādītas 5.1. tabulā. Runas korpusa gadījumā papildus specifiskajiem darbiem (līmeņiem) ir spēkā arī visas tekstu korpusa izstrādes.

Tekstu korpuss	
4. līmenis	Semantiskā meklēšana
3. līmenis	Sintaktiskie koki, sintaktiskā statistika
	Automatizētas apstrādes un analīzes rīki 4. līmeņa marķējuma iegūšanai
2. līmenis	Meklēšana, izmantojot morfoloģisko analīzi, statistika
	Automatizētas apstrādes un analīzes rīki 3. līmeņa marķējuma iegūšanai
1. līmenis	Navigācija, kontekstu ierobežošana un pozicionēšana, transformēšana prezentācijas formātos; vienkāršota meklēšana un statistika
	Automatizētas apstrādes un analīzes rīki 2. līmeņa marķējuma iegūšanai
0. līmenis	Automatizētas apstrādes un analīzes rīki 1. līmeņa marķējuma iegūšanai
-1. līmenis (ieguve)	Skenējot, savācot automātiski ar t. s. tīmekļa „zirnekli", manuāla ievade/rediģēšana

Runas korpuss	
2. līmenis	Runas sintēze
1. līmenis	Meklēšana (piem., jautājuma teikumu vai kāda runātāja atlase), intonāciju pētīšana u. c.
	Daļēji automātiska fonētiskā transkripcija
0. līmenis	Daļēji automātiska transkribēšana un prosodijas marķēšana
-1. līmenis (ieguve)	Audio datņu ieskaņošana (intervijas u. c.); periodiskas lejupielādes no atbilstošām tīmekļa vietnēm (piem., raidstacijas)

5.1. tabula – programmrīku izstrādes pa līmeņiem.



5.1. diagramma – korpusa sistēmas vienkāršots konceptuālais modelis.

6. Autortiesību (un autoratlīdzības) jautājums. Iespējamie risinājumi

6.1. Problēmas izklāsts (Ievads)

Valodas korpusu lietojums pasaulē pēdējo gadu laikā ir būtiski pieaudzis. Attiecīgi ir pieaudzis arī šo korpusu lietošanas spektrs, sākot no leksikogrāfijas un beidzot ar dažādām valodu tehnoloģijām. Paralēli ir attīstījusies arī valodas korpusu pieejas nodrošināšana tiešsaistes režīmā. Internets kļūst arvien plašāk pieejams, un internetā izvietotajai informācijai var piekļūt plašs lietotāju loks ne tikai vienas valsts ietvaros, bet arī globāli. Internetā izvietotajai informācijai potenciāli var piekļūt praktiski jebkurš pasaules iedzīvotājs neatkarīgi no viņa valstiskās piederības, dzīvesvietas un mērķa, kādam tas plāno izmantot korpusu.

Strādājot pie tiešsaistes valodas korpusa izveides, uzmanība ir jāpievērš ne tikai tehnoloģiskajiem risinājumiem, bet arī juridiska rakstura jautājumiem, lai varētu aizsargāt gan korpusā ievietoto resursu autoru, gan korpusa izstrādātāju, gan korpusa lietotāju likumiskās intereses. Turklāt bieži vien vērā ņemamie juridiskie aspekti izvirza īpašas prasības tehnoloģiskajiem risinājumiem.

Veidojot plašu un pilnvērtīgu korpusu latviešu valodai, praktiski nebūs iespējams izvairīties no aizsargātu darbu (t. i., autortiesību objektu) iekļaušanas, jo autoru darbi, t. sk. fonogrammas, ir būtiska korpusa sastāvdaļa. Līdz ar to juridiski regulējamie jautājumi galvenokārt ir saistīti ar autortiesību problēmām:

- darbu izmantošanas iespējām,
- autortiesību apjomu,
- teritorijas principu,
- autoratlīdzību, tās apmēru,
- citu valstu regulējumu autortiesību jomā u. c.

Savukārt intelektuālais īpašums jāaplūko no šādiem aspektiem:

- darbu autoru autortiesības – jāaplūko atšķirīgais regulējums gan vietējiem, gan ārzemju darbiem un autoriem,
- datu bāzes izstrādātāju autortiesības,
- valodas speciālistu intelektuālās tiesības uz izstrādāto risinājumu,
- tehnoloģiskā risinājuma izstrādātāju autortiesības,
- cita veida intelektuālais īpašums, piemēram, patenti.

Tā kā latviešu valodas korpusā paredzēts izvietot atšķirīgu autoru dažāda veida darbus gan latviešu, gan arī citās valodās (piemēram, paralēlajos korpusos), ir jāpiedāvā pietiekami universāli autortiesību regulēšanas risinājumi, kas būtu piemērojami katram darba veidam, ņemot vērā nianses, kas rodas sakarā ar darbu autoru valstisko piederību.

Autortiesības regulējošās normas, kas pastāv starptautiskā līmenī, mēģina harmonizēt atšķirīgās tiesību sistēmas un interpretācijas, taču tās sava starptautiskā rakstura dēļ nespēj nodrošināt precīzu un vienveidīgu regulējumu visās valstīs, bet tikai nosaka galvenās vadlīnijas, kas valstīm jāievēro. Starptautiskie līgumi, kas attiecas uz intelektuālo īpašumu, piedāvā valstīm pietiekamu elastību, ieviešot šo līgumu noteikumus nacionālajos tiesību aktos, jo valstis var ņemt vērā vietējos sociālos, kultūras un ekonomiskos apstākļus. Tomēr vienlaicīgi valstīm ir jāievēro noteikti minimālie standarti, par kuriem ir panāktas starptautiskas vienošanās. Intelektuālā īpašuma tiesības pēc savas būtības ir teritoriālas – piemēram, autortiesību likums, ko īsteno kādas valsts valdība, nosaka autortiesību prasības un nosacījumus, saturu un realizēšanu šajā valstī. Intelektuālā īpašuma likumu teritorialitāte rada problēmas tajā brīdī, kad intelektuālais īpašums ir pieejams un tiek lietots globālā telpā, piemēram, izmantojot internetu[90].

Tātad jebkura veida autoru darbu aizsardzībai var tikt piemērotas ne tikai viena likuma dažādas normas, bet pat pilnīgi cits likums vai pat piemērojamā tiesību sistēma. Tas ir izskaidrojams ar to, ka var atšķirties valsts, kuras normatīvie akti attiecināmi uz konkrēto darbu. Piemēram, ja darbs latviešu valodā izziņots Francijā un autors, kurš darbu radījis, nav piederīgs Latvijas Republikai, tad autora tiesības aizsargā Francijas likumi. Līdzīgi tiešsaistes resursam globālajā interneta tīklā var pieslēgties lietotāji no dažādām pasaules valstīm, tāpēc kļūst aktuāls jautājums par šādām attiecībām piemērojamo tiesību normu izvēli.

Lai nepārkāptu autoru tiesības un atvieglotu korpusa uzturētāju un lietotāju darbu, tiesiskās attiecības, kas rodas, izmantojot autoru darbus, ir ieteicams regulēt ar līgumu palīdzību. Šim mērķim tiesiskās attiecības var iedalīt divās lielās grupās:

1) attiecības starp korpusa uzturētājiem (īpašniekiem) un darbu autoriem – šajā gadījumā līgumi tiek slēgti ar darbu autoriem vai to pārstāvjiem;
2) attiecības starp korpusa uzturētājiem (īpašniekiem) un korpusa lietotājiem – līgumi tiek slēgti ar korpusa lietotājiem. Šajā gadījumā sākotnēji ir jādefinē, kas ir un kas var būt korpusa lietotāji. Tālākajos soļos ir jānosaka, kādā veidā lietotāji piekļūst korpusam, kādas darbības viņi drīkst veikt ar korpusu, kādam mērķim viņi drīkst izmantot korpusu. Ja korpuss tiek lietots par samaksu, arī par šo jautājumu ir jāvienojas līgumā.

Iniciatīva jāuzņemas korpusa uzturētājam, kas darbojas kā sava veida starpnieks starp autoriem un korpusa lietotājiem.

Tiesību akti, kas regulē autortiesību jomu, paredz dažādus izņēmumus, kas pieļauj ar autortiesībām aizsargātu darbu lietošanu, nemaksājot par to autoratlīdzību. Tas gan

[90] http://www.wipo.org – skatīts 01.07.2005.

nenozīmē, ka atbilstošs lietojums automātiski izslēdz autortiesību pārkāpumu, tāpēc līgumos ir jāiekļauj aspekti, kas atvieglo darbu izmantošanu atsevišķiem mērķiem.

Lai novērstu latviešu valodas korpusa izveides laikā apkopoto autoru darbu autortiesību pārkāpumus, kas saistīti gan ar atšķirīgajiem valstu likumiem, gan ar internetu kā korpusa vidi, būtu jāslēdz līgumi ar attiecīgo darbu autoriem. Pareizu tiesisko attiecību veidošana, kas ir pamats Latviešu valodas korpusa darbībai atbilstoši likumam, jāsāk līdz ar korpusa izstrādes uzsākšanu. Turpmāk tiek apskatīti tie svarīgākie jautājumi, kas jāievēro, gan veidojot Latviešu valodas korpusa datu bāzes sistēmu, programmatūru, gan darbu atlases, gan līgumu sagatavošanas un slēgšanas stadijās.

6.2. Lietotie termini

WIPO – *World Intellectual Property organization* (Pasaules Intelektuālā īpašuma organizācija).

Autortiesību objekts – likumā noteiktie autoru darbi neatkarīgi no to izpausmes veida un formas.

Autortiesību subjekts – darbu autori un līdzautori, un atvasināto darbu autori, kā arī minēto personu mantinieki un tiesību pārņēmēji.

***sui generis* tiesības** – īpaša veida, unikālas tiesības.

6.3. Tiesiskais regulējums atkarībā no korpusa lietojuma mērķa

Ir svarīgi izvērtēt mērķi, kādam tiks lietots valodas korpuss, kā arī definēt korpusa resursu pieejamību. Atkarībā no tā, vai tiek ņemta maksa par korpusa lietošanu, tiek izdalīti šādi tekstu korpusi:

1) **komerciālie** – par šo korpusu lietošanu jāmaksā konkrēta summa par noteiktu laika periodu. Komerciālo korpusu gadījumā var tikt piesaistīts iespējami plašs lietotāju loks, kā arī var brīvi noteikt korpusa izmantošanas mērķus un ļaut lietotājiem iegūt un izmantot pašus darbus. Šajā gadījumā summai ir jābūt pietiekamai, lai varētu segt darbu lietošanas izmaksas (autortiesību atlīdzība un korpusa uzturēšanas izmaksas);
2) **nekomerciālie** – resursu lietošana ir bez maksas. Šādi korpusi parasti ir pieejami slēgtam lietotāju lokam, piemēram, vienas institūcijas darbiniekiem vai pētniekiem.

Nekomerciāli korpusi ir tādi korpusi, kas tiek izmantoti tikai un vienīgi akadēmiskiem, t. i., pētnieciskiem mērķiem. Nekomerciāls korpuss tiek veidots ne kā literārs darbs vai šādu darbu apkopojums, ne arī kā faktoloģisks materiāls. Teksta korpusa izveides pamatmērķi var iedalīt divās lielās grupās, no kurām katrai ir savas lietojuma īpatnības:

– liela apjoma tekstu statistiska analīze – šajā gadījumā ir svarīgs nevis pats teksts, bet gan dažādas šī teksta īpašības, precīzāk, fakti par šo tekstu;
– leksisko īpašību pētīšana uz nelielu teksta fragmentu vai citātu bāzes – korpusa pētnieku vairāk interesē nevis teksta literārā vērtība, bet gan kāds konkrēts šī teksta aspekts no leksiskā viedokļa, t. i., korpusa lietotājam nav svarīgs teksta literārais vai cita veida konteksts.

Autortiesību likuma 21. pants pieļauj darbu izmantošanu zinātniskiem un pētniecības mērķiem. Nav iespējams viennozīmīgi noteikt, kāda veida zinātniska un pētnieciska darbība būtu atbilstoša 21. panta izpratnei, tomēr var secināt, ka noteiktas zinātniskas un pētnieciskas darbības iekļaujamas *bona fide* (laba ticība) lietošanas praksē. Eiropas tiesībās godprātīga darbu izmantošana ietver sevī kopēšanu personiskiem, zinātniskiem, izglītības vai citiem privātlietošanas mērķiem [Eisenchit, Turner 1997:209–223]. *Bona fide* lietojums ir atrunāts Bernes konvencijas 10. pantā, kā arī Latvijas Autortiesību likuma 20. un 21. pantā.

Autortiesību likuma 21. pants būtu vērtējams kopsakarā ar likuma 18. un 19. pantu. Proti, uz Latviešu valodas korpusu var attiecināt tikai 21. pantā noteikto, ka darba izmantošana izglītības un pētniecības mērķiem pieļaujama tikai līdzekļos, kas tiek speciāli radīti un izmantoti izglītības un pētniecības iestādēs nepastarpinātā mācību un pētniecības procesā to darbības mērķim atbilstošā apjomā nekomerciālos nolūkos. Tātad likums izvirza vairākas prasības, lai darbu varētu izmantot bez autora atļaujas un atlīdzības:

– mērķis – darba izmantošana izglītībā un pētniecībā;
– izglītības un pētniecības līdzekļi tiek radīti un izmantoti izglītības un pētniecības iestādēs;
– šos līdzekļus drīkst izmantot tikai nepastarpinātam mācību un pētniecības procesam;
– šos līdzekļus drīkst izmantot tikai un vienīgi nekomerciālā nolūkā.

Autortiesību likuma 18. pants nosaka svarīgu principu, kas jāievēro, veidojot Latviešu valodas korpusu izglītības un pētniecības mērķiem, noteiktie autora mantisko tiesību ierobežojumi piemērojami tikai tādā veidā, lai tie nebūtu pretrunā ar autora darba normālas izmantošanas noteikumiem un nepamatoti neierobežotu autora likumīgās intereses. Tas nozīmē, ka Latviešu valodas korpusa izveides rezultātā ir jārada pēc iespējas drošāka datu bāze, kas nepieļauj darbu kopēšanu. Autortiesību likuma 19. pants tikai nosaka, ka autortiesības netiek uzskatītas par pārkāptām, ja tiek ievērotas attiecīgajos pantos, kur paredzēti izņēmumi, noteiktās prasības.

Ja sākotnēji līgums par darba izmantošanu ar autoru noslēgts tikai izglītības un pētniecības mērķiem, tad vēlāk, kad Latviešu valodas korpuss kļūtu par komerciālu, visi

darbi, kas izmantojami tikai izglītības vai zinātnes mērķiem, būtu jāizņem no Latviešu valodas korpusa datu bāzes, jo neatbilstu datu bāzes komerciālajam mērķim. Tāpēc pastāv divi risinājumi: vai nu Latviešu valodas korpusa koncepcijā paredzēt noteiktu korpusa attīstības plānu, paredzot iespēju, ka korpuss var būt arī komerciāls, vai arī slēdzot tādus līgumus ar autoriem, kas paredz iespēju saņemt autoratlīdzību, ja Latviešu valodas korpuss kļūs komerciāls.

6.4. Tekstu uzglabāšana un lietošana

Teksti parasti tiek uzglabāti trīs formātos:

- oriģinālie faili – pilni teksti,
- teksti HTML formātā – iespējama pilna tekstu apskate,
- ASCII faili, kas ir pamatā HTML failiem – tiek izmantoti teksta lingvistiskajā analīzē.

Teksta korpusu izmantošanā pētnieki var piekļūt šo failu kopijām, kā arī atkarībā no teksta korpusa piedāvātajām iespējām veikt izmaiņas šajos tekstos. Pieļaujamās darbības ir iekļaujamas teksta korpusa lietošanas līgumā:

1) teksta korpusa lietošana – izmantošanas apjoms un mērķis, lietošanas noteikumi,
2) tekstu papildināšana.

Ja lietotājs var tekstu papildināt, mainīt vai veikt tamlīdzīgu tā apstrādi, kas tam nepieciešama pētniecības procesā, saskaņā ar Autortiesību likuma 14. p. 5. daļu, kurā uzskaitītas darbu autoru personiskās tiesības, tiesības saistībā ar darba pārveidošanu ir jāprasa darba autoram, slēdzot līgumu par darba iekļaušanu tekstu korpusā. Šajā gadījumā ir arī jāņem vērā tā paša panta 6. daļā paredzētās autora tiesības vienpusēji atkāpties no līguma, ja:

1) darbs tiek izkropļots, pārveidots vai citādi sagrozīts,
2) autora tiesību aizskārums var kaitēt autora godam un cieņai.

Tomēr jāņem vērā arī tas, ka teksta pārveidošana tiek veikta pētnieka personiskajām vajadzībām un pārveidotie teksti parasti netiek publiskoti. Tāpēc, lai izvairītos no autoru neapmierinātības, pārveidotu tekstu publiska pieejamība jāizslēdz.

Ja teksta korpusa funkcionalitāte pieļauj iespēju iegūt kāda darba pilnu tekstu, jācenšas modelēt tāda situācija, kurā autortiesību aizsardzība tekstiem būtu minimāla. To var panākt:

1) izmantojot tekstus, kuriem notecējis autortiesību aizsardzības termiņš;
2) iekļaujot korpusā tekstu fragmentus, nevis pilnas versijas;
3) ierobežojot un kontrolējot lietotāju pieejas tiesības;

4) nodrošinot, ka izejas teksts būtiski atšķiras no sākotnējā, ar autortiesībām aizsargātā teksta;

5) atrunājot teksta lietošanu līgumos ar autoriem.

6.5. Intelektuālo tiesību apjoms

Latvijas Autortiesību likuma 3. pants nosaka to autoru darbu loku, uz ko attiecināms Latvijas Republikas normatīvo aktu regulējums autortiesību jomā. Latvijas Autortiesību likums aizsargā darbus, kas:

- izziņoti Latvijā;
- nav izziņoti Latvijā, bet atrodas Latvijā jebkādā materializētā formā;
- vienlaikus publicēti ārvalstī un Latvijā;
- izziņoti ārvalstī jebkādā materializētā formā un kurus radījuši Latvijas pilsoņi vai personas, kurām ir tiesības uz nepilsoņa pasi, vai personas, kurām Latvija ir pastāvīgā dzīvesvieta (domicils).

Citu personu autortiesības uz darbiem, kas izziņoti vai citādi darīti zināmi ārvalstī jebkādā materializētā formā, tiek atzītas saskaņā ar Latvijai saistošiem starptautiskajiem līgumiem.

Eiropas Savienības ietvaros, lai harmonizētu atšķirīgās nacionālās tiesību normas, pastarpināti piemērojamie normatīvie akti – direktīvas.

Pasaules mērogā šo tiesību harmonizāciju nodrošina konvencijas intelektuālā īpašuma jomā. Tomēr vēl joprojām starp dažādām valstīm pastāv būtiskas atšķirības. Kā piemēru šeit var minēt ASV.

Atšķirīgs tiesību apjoms attiecas uz blakustiesību darbiem, jo svarīga ir ne tikai darbu saistība ar valsti, bet arī blakustiesību subjektu saistība ar valsti. Atšķirīgiem blakustiesību subjektiem, lai tiktu atzīta konkrētās valsts aizsardzība, piemērojami savādāki nosacījumi. Tā Latvijas Autortiesību likuma 56. pants nosaka, ka:

1. Izpildītāju tiesības tiek atzītas, ja pastāv kāds no šādiem nosacījumiem:

- izpildītājs ir Latvijas pilsonis vai persona, kurai ir tiesības uz Latvijas nepilsoņa pasi, vai persona, kurai Latvija ir pastāvīgā dzīvesvieta (domicils);
- izpildījums veikts Latvijā;
- izpildījums fiksēts fonogrammā, kas ir aizsargāta ar Latvijas Autortiesību likumu;
- izpildījums, kas nav fiksēts fonogrammā, iekļauts raidorganizācijas raidījumā, kuras oficiālā atrašanās vieta ir Latvija.

2. Fonogrammu producentu tiesības tiek atzītas, ja pastāv kāds no šādiem nosacījumiem:

- fonogrammas producents ir Latvijas pilsonis vai persona, kurai ir tiesības uz Latvijas nepilsoņa pasi, vai persona, kurai Latvija ir pastāvīgā dzīvesvieta (domicils);
- skaņas pirmā fiksācija veikta Latvijā;
- fonogrammas publicēšana vai publiskošana veikta Latvijā.

3. Raidorganizāciju tiesības saskaņā ar šo nodaļu tiek atzītas, ja raidorganizācijas oficiālā atrašanās vieta ir Latvija.

6.6. Latviešu valodas korpusā iekļaujamo darbu veidi

Latviešu valodas korpusā plānots izmantot materiālā formā fiksētus darbus. Autortiesību likums nosaka šādus rakstiskā formā fiksētus darbus, kas var tikt iekļauti Latviešu valodas korpusā: literārie darbi, dramatiskie un muzikāli dramatiskie darbi, scenāriji, audiovizuālo darbu literārie projekti, muzikālie darbi ar tekstu, atvasinātie darbi, fonogrammas, normatīvie un administratīvie akti, citi valsts un pašvaldību iestāžu izdoti dokumenti, tiesas nolēmumi, presē, radio vai televīzijas raidījumos vai citos saziņas līdzekļos sniegtā informācija par dienas jaunumiem, dažādiem faktiem un notikumiem. Šis uzskaitījums nav izsmeļošs, jo zem uzskaitītajiem darbiem iekļauti arī citi, piemēram, literārie darbi ietver arī zinātnisko un populārzinātnisko literatūru. Šis darbu uzskaitījums un dalījums ir svarīgs no juridiskā viedokļa, jo jebkurš darbs, ko vēlēsies ietvert Latviešu valodas korpusā, būs jāizvērtē tieši no Autortiesību likumā piedāvātās darbu klasifikācijas viedokļa.

Autortiesību likuma 1. panta 2. punkts definē jēdzienu „darbs" – autora radošās darbības rezultāts literatūras, zinātnes vai mākslas jomā neatkarīgi no tā izpausmes veida, formas un vērtības. Svarīgi norādīt, ka Latvijas likums paredz, ka aizsargājams ir tikai tāds darbs, kas izveidots radošas darbības rezultātā. Grūti pierādāms un apliecināms ir darba tapšanas procesa radošais moments, tāpēc, vadoties pēc autortiesību principa, ka šīs tiesības ir vērstas uz autoru interešu aizsardzību, tomēr jebkurš darbs, ko radījis autors, būtu uzskatāms par aizsargājamu. Šo apgalvojumu apstiprina arī asoc. prof. Jānis Rozenfelds: „Daži autortiesību speciālisti uzskata par pietiekamu norādi uz darbu, citi uzskata par nepieciešamu pievienot arī norādi par to, ka darbs ir jaunrades rezultāts. Šī formula var radīt papildu neskaidrības, jo it kā norāda uz nepieciešamību nodibināt autora radošās darbības pakāpi, lai noskaidrotu jautājumu, vai viņa darbu aizsargā autortiesības vai ne. Patiesībā šāda nepieciešamība nepastāv, un tādēļ norāde uz darbu kā jaunrades rezultātu šķiet pilnīgi lieka" [Rozenfelds 2004:25].

Plašāks jēdziena „darbs" skaidrojums rodams starptautiskajos tiesību aktos. 1886. gada 9. augusta Bernes konvencija par literatūras un mākslas darbu aizsardzību

2. pantā uzskaita aizsargājamos literāros darbus, turklāt radošā darbība tiek attiecināta tikai uz literatūras un mākslas darbu krājumiem (piemēram, enciklopēdijas, antoloģijas), kas materiālu atlases un izkārtojuma ziņā ir intelektuālās jaunrades rezultāts. 1996. gada 20. decembra WIPO autortiesību līgums 2. pantā norāda, ka autortiesību aizsardzība nav attiecināma uz idejām, procesiem, darbības metodēm vai matemātiskām koncepcijām kā tādām. Latvijas Autortiesību likumā ir iekļautas WIPO autortiesību līguma nostādnes. Latvijas likums neaizsargā:

- normatīvos un administratīvos aktus, citus valsts un pašvaldību iestāžu izdotus dokumentus un tiesas nolēmumus, kā arī šo tekstu konsolidētās versijas;
- presē, radio vai televīzijas raidījumos vai citos saziņas līdzekļos sniegto informāciju par dienas jaunumiem, dažādiem faktiem un notikumiem;
- idejas, metodes, procesus un matemātiskās koncepcijas.
- Attiecībā uz blakus tiesību objektiem (fonogrammas, fiksācijas) ir paredzēts atšķirīgs regulējums. Latvijas Autortiesību likums norāda, ka:
- fiksācija ir skaņas vai attēla iemiesojums materializētā formā, kas dod iespēju publiskot, uztvert vai reproducēt ar attiecīgas ierīces palīdzību;
- fonogramma ir izpildījuma skaņu, citu skaņu vai skaņu atveidojuma fiksācija.

Šo blakus tiesību objektu tiesību regulējums atrodams Latvijas Autortiesību likuma 8. nodaļā. 1996. gada 20. decembra WIPO līgums par izpildījumu un fonogrammām sniedz līdzīgu skaidrojumu. Atšķirības starp fiksāciju un fonogrammu:

- fiksācija ir plašāks jēdziens, kas iekļauj audiovizuālu darbu fiksāciju, fonogrammu fiksāciju vai tikai vizuālu darbu fiksāciju;
- fonogramma ir šaurāks jēdziens, kas attiecināms tikai uz skaņas fiksāciju.

6.6.1. *Tekstu korpuss*

Latviešu valodas korpusa tekstu daļā var iekļaut gan ar autortiesībām aizsargājamus, gan neaizsargājamus darbus latviešu valodā, bet, ja tiek veidots paralēlais valodas korpuss, – arī citās valodās. Latvijā par neaizsargājamiem uzskatāmi koncepcijas 6. 6. punktā norādītie darbi. Tas, ka darbs netiek aizsargāts, nozīmē, ka nav nepieciešama autora atļauja tālākai darbu izmantošanai, kā arī nav jāmaksā autoratlīdzība. Citās valstīs neaizsargājamo darbu uzskaitījums var atšķirties.

Latvijas Autortiesību likums uzskaita aizsargājamos darbus 4. pantā. Tekstu korpusam no uzskaitītajiem ir piederīgi:

- literārie darbi,
- dramatiskie darbi,

– muzikāli dramatiskie darbi,
– scenāriji,
– muzikālie darbi ar tekstu,
– atvasinātie darbi.

Lai izmantotu uzskaitītos darbus, ir nepieciešama autora atļauja, kā arī par darba izmantošanu autoram pienākas autora atlīdzība. Autortiesību likums paredz arī izņēmuma gadījumus, kad nav nepieciešama autora atļauja un autoratlīdzības samaksa (sk. koncepcijas 6. 8. punktu).

Literāri darbi no autortiesību viedokļa ir jebkuri oriģināldarbu teksti – vai tā būtu daiļliteratūra, vai tehniski, zinātniski, praktiskas rokasgrāmatas[91]. Latvijas autortiesību likums uzskaita šādus literāros darbus:

– grāmatas;
– brošūras;
– runas;
– lekcijas;
– aicinājumus;
– ziņojumus;
– spredikļus.

Taču seko piebilde „un citi līdzīga veida darbi". Likumā arī minēti nepabeigti darbi, kas līdzīgi kā citi darbi ir autortiesību objekts. 1886. gada 9. augusta Bernes konvencija par literatūras un mākslas darbu aizsardzību norāda, ka valstis var atzīt par aizsargājamiem likumdošanas, administratīva un juridiska rakstura oficiālos tekstus un šādu tekstu oficiālos tulkojumus. Latvijas Autortiesību likums norāda, ka normatīvie un administratīvie akti, citi valsts un pašvaldību iestāžu izdotie dokumenti un tiesas nolēmumi, kā arī šo tekstu konsolidētās versijas netiek aizsargātas. Taču gadījumos, kas attiecas uz citu valstu normatīvo aktu tulkojumiem, ir jābūt piesardzīgiem un jāpārbauda izmantošanas nosacījumi.

Dramatiskie darbi – monologa vai biežāk dialogu formā rakstīts literārs darbs, savstarpēji saistītu darbību un diskursu kompilācija, kas personas atklāj darbībā skatuviskā vai tamlīdzīgā iestudējumā[92]. Kas attiecas uz dramatiskajiem darbiem, tad valodas korpusā var izmantot rakstiskā formā fiksēto materiālu, kas tiek aizsargāts tāpat kā jebkurš literārs darbs.

Muzikāli dramatiskie darbi – dramatisks darbs, kura būtiska un neatņemama sastāvdaļa ir mūzika. Muzikāli dramatisks darbs ir opera, operete, balets, mūzikls.

[91] http://www.autornet.lv/tiesibas/vardnica – skatīts 10.07.2005.
[92] http://www.autornet.lv/tiesibas/vardnica – skatīts 28.06.2005.

Mūzika parasti šajos darbos saplūst ar libretu, veidojot vienotu veselumu[93]. Līdzīgi kā dramatiskajiem darbiem arī muzikāli dramatiskajiem darbiem valodu korpusā pielietojama tikai rakstiskā formā fiksētā daļa, turklāt tikai libreta daļa – tas ir literārs darbs, kas rakstīts muzikāli dramatiska darba radīšanai[94].

Scenārijs – parasti ir audiovizuāla darba (arī skatuviskas darbības vai radio un televīzijas raidījuma) teksts rakstiskā formā. To var publicēt atsevišķi tāpat kā jebkuru citu literāru darbu[95].

Muzikāls darbs – ir visdažādāko skaņu kombinācijas (kompozīcijas) ar vai bez teksta (dzejas teksta), kas radīts izpildīšanai ar mūzikas instrumentiem un/vai ar balsi[96]. Valodas korpusā iespējams izmantot tekstuālo daļu, kas attiecas uz dziesmu vārdiem (dzejas teksts).

Atsevišķi ir jāapskata tādu darbu autortiesību jautājumi kā:

- publikācijas plašsaziņas līdzekļos;
- publikācijas internetā.

Saskaņā ar šobrīd spēkā esošajos Latvijas likumos noteikto kārtību interneta vide netiek pielīdzināta plašsaziņas līdzekļiem. Autortiesību likuma 6. pants cita starpā min, ka ar autortiesībām netiek aizsargāta presē, radio vai televīzijas raidījumos vai citos saziņas līdzekļos sniegtā informācija par dienas jaunumiem, dažādiem faktiem un notikumiem. Šī definīcija aptver gan plašsaziņas līdzekļus, gan interneta vidi. Pārējie darbi, kas tiek publiskoti, ir aizsargāti ar Autortiesību likumu un tiek pieskaitīti pie literārajiem darbiem.

Taču jābūt piesardzīgiem arī attiecībā uz rakstiem, kas it kā pēc pazīmēm atbilst Autortiesību likuma 6. panta 4. punktam un tāpēc netiek aizsargāti. Bieži vien darbs var saturēt tikai apkopotu informāciju, tomēr šī darba autoram tas ir prasījis daudz laika un pietiekami lielas izmaksas, tāpēc šādi darbi var tikt aizsargāti ar konkurenci regulējošajām normām, jo rezultāts ir „produkts", kas tiek piedāvāts pircējiem.

Atvasinātie darbi – darbs, kas radīts izmantojot citu oriģināldarbu, turklāt jaunais darbs ir atvasināts no oriģināldarba. Atvasinātie darbi ir aizsargāti neatkarīgi no tā, vai darbi, uz kuriem atvasinātie darbi balstās vai kuri iekļauti tajos, ir aizsargāti darbi vai ne[97].

Atvasinātie darbi, kas var tikt iekļauti Latviešu valodas korpusā, var būt:

- tulkojumi,
- atlanti,

[93] Ibid.
[94] Ibid.
[95] Ibid.
[96] Ibid.
[97] Ibid.

– enciklopēdijas,
– antoloģijas,
– hrestomātijas,
– referāti,
– kopsavilkumi,
– apskati,
– darbu krājumi,
– citi salikti darbi.

Apkopojot iepriekš uzskaitītos darbu veidus, var secināt, ka teksta korpusā var iekļaut jebkuru rakstiskā formā izteiktu informāciju. Pirms ievietošanas Latviešu valodas korpusā ir jāpārbauda, vai attiecīgais teksts ir pakļauts autortiesību aizsardzībai. Ja teksts ir pakļauts autortiesību aizsardzībai, tad jārīkojas **saskaņā ar koncepcijas 6. punktā zemāk piedāvātajiem risinājumiem.**

6.6.2. *Runas korpuss*

Runa – literārs darbs, kas radīts specifiskā veidā izziņošanai runas jeb mutvārdu formā. To nevajag sajaukt ar darbu, kas radīts uztverei ar redzi (publicēšanai), bet izziņots, to publiski nolasot (uztverei ar dzirdi). Runas ir: lekcija, aizstāvības runa tiesā, politiska darbinieka runa, sprediķis u. c.[98] Runa var būt uzrakstīta uz papīra, tādējādi kļūstot par literāru darbu, kas tiek aizsargāts ar autortiesībām, taču gadījumos, kad runa tiek fiksēta fonogrammā, pastāv ne tikai runas autora tiesības, bet arī izpildītāja un fonogrammas producenta tiesības, ko dēvē par blakustiesībām.

Fonogramma ir muzikālu vai jebkuru citu skaņu ekskluzīvi aurālas fiksācijas (tādas, ko var uztvert ar dzirdi) rezultāts magnētiskajās lentēs, disketēs vai optiskajās ierīcēs (vienalga, kādā veidā – digitālā vai analogā). Skaņuplates, audiokasetes, CD u. c. ir fonogrammu reproducēšanas rezultāts – fonogrammu dublikāti.[99]

Runas korpusā paredzēts izvietot runātus darbus bez muzikāla pavadījuma vai noformējuma. Tie varētu būt monologi vai dialogi, nesagatavota vai sagatavota runa. Pastāv iespēja izmantot raidorganizāciju sagatavotos ierakstus (gan radio, gan televīzijas). Lai iekļautu šos darbus Latviešu valodas runas korpusa daļā, ir nepieciešams saņemt atļauju no attiecīgā uzņēmuma, kas radījis raidījumu.

Jāatzīmē, ka lielākoties esošie runu korpusi tiek izmantoti mācību un pētniecības nolūkos un ir nekomerciāli. Kā piemēru var minēt plašo krājumu, kas atrodas mājas lapā www.dzivesstasts.lv. Šāda veida savākti ieraksti, kas paredzēti mācību un pētniecības nolūkiem, nedrīkst tikt izmantoti citiem mērķiem, jo, izmantojot materiālus citam

[98] Ibid.
[99] Ibid.

nolūkam, tiktu pārkāptas izpildītāju un ierakstu izveidotāju tiesības, kas tiek aizsargātas ar Autortiesību likumu. Kā norādīts mājas lapā www.dzivesstasts.lv: „Dzīvesstāstu krāšana notiek pētījumos un ekspedīcijās. Dzīvesstāsta intervija ir autora – stāstītāja un intervētāja – stāsta virzītāja, sadarbības process un rezultāts. Autors par savu dzīvi stāsta pats. Bet ierosinājis stāstīt ir otrs cilvēks – intervētājs. Intervētāja uzdevums ir ierosināt sarunu, kas ir brīva no kvantitatīvo pētījumu vai strukturēto aptauju piedāvātajiem standartiem. Intervētājs virza sarunu neformāli, lai sarunu biedrs par pierastām situācijām un savas dzīves notikumiem var stāstīt ikdienas sarunu valodā." Intervētājs faktiski ir fiksācijas veicējs, kuram pieder fonogrammu producentam noteiktās tiesības, savukārt stāstītājs ir gan autors, gan izpildītājs. Šī runas korpusa mērķis ir ne tikai sarunu valodas pētniecība, bet arī vēstures notikumu atspoguļošana. Dzīvesstāstu krājums sastāv arī no skaņu (runas) fonogrammām, ko ierakstījuši žurnālisti. Autori jeb stāstītāji ir devuši atļauju izmantot stāstījumu pētniecības un izglītības mērķiem, kā norādīts www.dzivesstasts.lv. Savukārt attiecībā uz fonogrammu producentiem jeb intervētājiem ir noteikts, ka tiem pieder darba pirmpublikācijas tiesības. Lietošanas noteikumi detalizēti nosaka mērķus, kuru sasniegšanai drīkst izmantot piedāvātos avotus, proti, zinātniskiem mērķiem, studentu zinātniskiem darbiem, bakalauru, maģistru un doktorantūras pētījumiem, novadu kultūras un identitātes pētījumiem, iemaņu, tradīciju, paradumu, saziņas formu apzināšanai, mācībām, izglītības programmām, grāmatu, filmu un raidījumu veidošanai.

6.7. Autora un citu autortiesību subjektu tiesības

Autortiesību mērķis ir autora tiesību aizsardzība. Latviešu valodas korpusā iekļaujamo darbu apkopošanas laikā būs nepieciešams iegūt darbu izmantošanas atļaujas. Darba izmantošanas atļaujas noteiktās situācijās var sniegt atšķirīgi autortiesību un blakustiesību subjekti. Tā kā autora intereses var pārstāvēt arī citas personas, tad autortiesību regulējums tiek attiecināts arī uz šiem subjektiem. Jāatzīmē, ka gadījumos, kad autors ir rīcībnespējīgs (autors ir nepilngadīgais vecumā līdz 18 gadiem vai ar tiesas spriedumu par rīcībnespējīgu atzīta persona), tad šo personu intereses pārstāv attiecīgi vai nu aizgādņi, vai aizbildņi, ievērojot Latvijas Republikas Civillikuma normas. Autortiesību likuma 7. pants nosaka to peronu loku, kuras var tikt atzītas par autortiesību subjektiem: pats autors, līdzautori, atvasināto darbu autori, mantinieki, kā arī citi autortiesību pārņēmēji.

Autoram pieder:

- personiskās tiesības jeb nemantiskās tiesības;
- mantiskās tiesības.

Tas nozīmē, ka autortiesības sastāv no atšķirīgu tiesību kopuma. Personiskās tiesības ir autoram neatsavināmas un pieder autoram visu viņa dzīves laiku un vismaz vēl tik

ilgi, kamēr spēkā ir autortiesību aizsardzības noteiktais termiņš pēc autora nāves. Personiskās tiesības ir neatsavināmas arī gadījumā, kad autors devis savu piekrišanu šo tiesību atsavināšanai.

6.7.1. *Autora personiskās tiesības*

Personiskās tiesības ir tiesības uz:

1) **autorību** – tiesības tikt atzītam par autoru;
2) **izlemšanu**, vai darbs tiks izziņots un kad tas tiks izziņots;
3) **darba atsaukšanu** – tiesības pieprasīt, lai darba izmantošana tiek pārtraukta, ar noteikumu, ka autors sedz zaudējumus, kas pārtraukšanas dēļ radušies izmantotājam;
4) **vārdu** – tiesības pieprasīt, lai viņa vārds būtu pienācīgi norādīts visās kopijās un jebkurā ar viņa darbu saistītā publiskā pasākumā, vai pieprasīt pseidonīma lietošanu vai anonimitāti;
5) **darba neaizskaramību** – tiesības atļaut vai aizliegt izdarīt jebkādus pārveidojumus, grozījumus un papildinājumus gan pašā darbā, gan tā nosaukumā;
6) **pretdarbību** (arī uz vienpusēju atkāpšanos no līguma, neatlīdzinot zaudējumus) jebkurai sava darba izkropļošanai, sagrozīšanai vai citādai pārveidošanai, kā arī tādai autora tiesību aizskaršanai, kas var kaitēt autora godam vai cieņai.

Izvietojot darbus Latviešu valodas korpusā, būtu jānodrošina autora un blakustiesību subjektu personiskās tiesības.

Tiesības uz autorību. Tiesības uz autorību un vārdu ir noteiktas Bernes konvencijas 6. bis pantā. Autoram ir tiesības pieprasīt, lai tiktu atzīta viņa autorība uz darbu, kā arī pie katras autora darba izmantošanas uz katra darba eksemplāra būtu norādīts darba autors. Šis pants arī norāda, ka autors var iebilst pret jebkādu šā darba sagrozīšanu, izkropļošanu vai citādu pārveidošanu, kā arī pret jebkādu citādu darba autora tiesību aizskaršanu, kura var kaitēt autora godam vai reputācijai. Ir valstis, kur kā papildu prasība tiek noteikts, ka jānorāda autora tituls vai cits goda nosaukums[100].

Darba neaizskaramība. Darba neaizskaramība, kā norādīts Autortiesību likumā, ir aizliegums veikt jebkādus pārveidojumus darbā. Par pretlikumīgu tiek uzskatīta jebkura rīcība, kas ir saistīta ar atkārtotu darba publicēšanu bez autora atļaujas, tieša darba pārveidošana, tai skaitā mainot darba saturu. Jebkuras izmaiņas autora darbā ir jāsaskaņo ar darba autoru. Bernes konvencijā darba neaizskaramības princips valstu likumos tiek traktēta divos veidos.

[100] http://www.lpia.lv/?id=373 – skatīts 05.07.2005.

1. Pirmajā jeb dinamiskajā koncepcijā norma tiek pārrakstīta ar tādu pašu jēgu kā konvencijā, tas ir, autoram tiek dota likumiska iespēja aizliegt tikai tādu darba pārveidošanu, kas var nodarīt kaitējumu autora cieņai vai reputācijai. To, vai konkrētajā gadījumā minētais kaitējums ir vai nav cieņu vai reputāciju aizskarošs, izvērtē tiesa.
2. Otrā jeb statiskā koncepcija aizliedz jebkuru darba modifikāciju, nedodot iespēju tiesai lemt par to, vai ir noticis cieņas vai reputācijas aizskārums [101].

Atkarībā no tā, no kuras koncepcijas vadījies konkrētās valsts likumdevējs, ir jāizvērtē izdevēja rīcība, pārveidojot darbu. Pēc statiskās koncepcijas loģikas, izdevējs nav tiesīgs izdarīt nekādus labojumus darbā, tomēr gan jurisprudence, gan doktrīna pieļauj izmaiņas darbā, kas saistītas ar gramatikas, lingvistikas un citiem tamlīdzīgiem labojumiem[102].

Citas personiskās tiesības. Pārējās Autortiesību likuma 14. pantā uzskaitītās autora personisko tiesību normas saistītas ar atkāpšanos no līguma vai aizlieguma publiskot darbu. Attiecībā uz izziņošanas aizliegumu pastāv salīdzinoši mazs risks, ka tiks radīts materiāls zaudējums citām personām, izņemot pašu autoru. Taču attiecībā uz darba atsaukšanu un pretdarbību pastāv liels risks, ka šāda darbība radīs zaudējumus līgumslēdzējiem. Vienā gadījumā autoram ir pienākums atlīdzināt radītos zaudējumus, taču otrā darbība ir paredzēta savu tiesību aizsardzības nolūkos un tāpēc zaudējumi nav jāatlīdzina.

Autoru personisko tiesību ievērošana korpusā. Lai ievērotu autora tiesības, nepieciešams Latviešu valodas korpusā pie katra darba vai tā fragmenta norādīt autora vārdu, uzvārdu, darba nosaukumu. Cik izsmeļošai ir jābūt informācijai atsaucē uz autora vārdu? Autortiesību likums norāda minimālās prasības, kas ir pietiekamas, lai identificētu darba autoru, proti, jānorāda autora vārds, uzvārds. Ja darbs tiek citēts vai tiek publiskots darba fragments, jānorāda arī darba nosaukums. Ikdienā sastopami atšķirīgu atsauču saturi, kurās tiek iekļauts autora vārds, uzvārds, darba nosaukums, izdevēja nosaukums, izdošanas gads un darba lapaspuses numurs, no kura ņemts fragments vai citāts. Šādu datu norādīšana atvieglo autortiesību atbilstības pārbaudi, bet tā nav obligāta prasība, izņemot gadījumus, ja tas noteikts ar iekšējo normatīvo aktu palīdzību kā obligāta prasība iestādes ietvaros.

6.7.2. *Autora mantiskās tiesības*

Autora mantiskās tiesības atšķirībā no personiskajām tiesībām ir atsavināmas un nododamas citiem autortiesību pārņēmējiem. Šīs tiesības nodrošina autoram atlīdzības saņemšanas iespēju par radīto darbu. Mantiskās tiesības ir autora tiesības izlemt:

[101] Ibid
[102] Ibid.

– publiskot darbu;
– publicēt darbu;
– publiski izpildīt darbu;
– izplatīt darbu;
– raidīt darbu;
– retranslēt darbu;
– padarīt darbu pieejamu sabiedrībai pa vadiem vai citādā veidā tādējādi, ka tam var piekļūt individuāli izraudzītā vietā un individuāli izraudzītā laikā;
– iznomāt, izīrēt vai publiski patapināt darba oriģinālu vai kopijas, izņemot trīsdimensiju arhitektūras darbus un lietišķās mākslas darbus;
– tieši vai netieši, īslaicīgi vai pastāvīgi reproducēt darbu;
– tulkot darbu;
– aranžēt, dramatizēt, ekranizēt vai citādi pārveidot darbu.

Šīs tiesības autors var atsavināt, turklāt autors var brīvi noteikt apjomu, termiņu un citus nosacījumus, kas saistīti ar darba atsavināšanu. Tā kā Latviešu valodas korpuss būs piesaistīts globālajam interneta tīklam, tad korpusa veidotājiem no izvēlētā darba autora ir jāiegūst atļauja padarīt darbu pieejamu sabiedrībai pa vadiem vai citādā veidā tādējādi, ka tam var piekļūt individuāli izraudzītā vietā un individuāli izraudzītā laikā. Latviešu valodas korpusa izdošana CD formātā būtu neizdevīga gan autoriem, gan korpusa veidotājiem, jo pastāv risks, ka tiek veidotas pirātiskās kopijas, kā arī ir grūtāk kontrolēt cita veida neatļautu darba izmantošanu. Ja kāda daļa vai pilns korpuss tomēr tiek izdots CD formātā, tehnoloģiskajam risinājumam ir jānodrošina tekstu aizsardzība tā, lai nebūtu iespējams (pēc iespējas apgrūtinot) teksta iegūšanu pilnā formā.

6.7.3. *Blakustiesības*

Latviešu valodas korpusa runas korpusā ievietotie darbi tiek aizsargāti ne tikai ar autortiesībām, bet arī blakustiesībām. Blakustiesību subjektu tiesību apjoms atšķiras no citu autortiesību subjektu tiesībām, jo radīto darbu raksturs un personu ieguldījums ir atšķirīgs. Blakustiesību aizsargāto darbu radīšanā piedalās vismaz divas personas. Protams, pastāv iespēja, ka gan izpildītājs, gan fonogrammu producents ir viena un tā pati persona, bet tādā gadījumā tiek dalīts personas ieguldījums gan kā izpildītājam, gan kā producentam. Pastāv šādas blakustiesību aizsargāta darba radīšanas shēmas:

1) darba autors ir radījis darbu un piesaista attiecīgi izpildītāju, kurš izpilda darbu tā, kā to vēlas darba autors, darba autors arī piesaista fonogrammas producentu, kas veic autora darba izpildījuma skaņu vai skaņu atveidojuma pirmo fiksāciju un ir atbildīgs par tās pabeigšanu;

2) fonogrammu producents ir saņēmis atļauju no darba autora par darba izmantošanu, piesaista darba izpildītāju un veic izpildījuma fiksāciju;
3) izpildītājs ir saņēmis atļauju no darba autora par darba izmantošanu, piesaista fonogrammu producentu, kurš veic izpildījuma fiksāciju;
4) izpildītājs ir darba autors un veic izpildījumu, ko fiksē fonogrammas producents;
5) izpildītājs ir darba autors un pats veic fonogrammu producenta pienākumus, veicot izpildījuma fiksāciju.

Tātad skaņu ierakstu gadījumā var konstatēt vismaz trīs dažādus tiesību apjomus, ko var realizēt viena vai vairākas personas. Darba autora tiesības tiek aizsargātas iepriekš aprakstītā autortiesību aizsardzības noteiktajā kartībā. Specifiskas ir izpildītāju un fonogrammu producentu tiesības. Izpildītājam pieder gan personiskās, gan mantiskās tiesības uz izpildījumu. Izpildītāja personiskās tiesības ir ierobežotas un ir atšķirīgas no darba autora tiesībām. Autortiesību likums nosaka, ka izpildītājam ir šādas personiskās tiesības neatkarīgi no viņa mantiskajām tiesībām, kā arī gadījumā, kad mantiskās tiesības tiek nodotas, attiecībā uz savu izpildījumu un tā fiksāciju ir tiesības:

– prasīt, lai izpildītājs tiktu identificēts, izņemot gadījumus, kad tas nav iespējams izpildījuma izmantošanas veida dēļ;
– iebilst pret jebkādu viņa izpildījuma izkropļošanu, sagrozīšanu vai citādu pārveidošanu, kas var kaitēt izpildītāja reputācijai.

Līdzīgi kā autoru personisko tiesību gadījumā arī izpildītāju personiskās tiesības nepāriet citām personām, kā arī nav atsavināmas pat tad, ja tam piekrīt pats izpildītājs.

Attiecībā uz izpildītāja mantiskajām tiesībām Autortiesību likums paredz šādu apjomu:

– raidīt vai publiskot izpildījumu, izņemot gadījumus, kad izpildījums ir jau raidīts;
– fiksēt agrāk nefiksētu izpildījumu;
– izplatīt izpildījuma fiksāciju;
– raidīt vai retranslēt pa kabeļiem izpildījuma fiksāciju;
– padarīt izpildījuma fiksāciju pieejamu sabiedrībai pa vadiem vai citādā veidā tādējādi, ka tai var piekļūt individuāli izraudzītā vietā un individuāli izraudzītā laikā;
– nomāt, īrēt un publiski patapināt izpildījuma fiksāciju;
– tieši vai netieši, īslaicīgi vai pastāvīgi reproducēt izpildījuma fiksāciju.

Arī runas korpusa daļas izveidošanai nepieciešams iegūt atļauju no blakustiesību subjektiem par darba izmantošanu, padarot izpildījuma fiksāciju pieejamu sabiedrībai pa

vadiem vai citādā veidā tādējādi, ka tai var piekļūt individuāli izraudzītā vietā un individuāli izraudzītā laikā. Tā kā blakustiesību objektu rada vairākas personas, tad gan no izpildītāja, gan no fonogrammas producenta būtu jāiegūst atļauja darbu izmantot. Starp izpildītāju un fonogrammas producentu noslēgts līgums var noteikt atšķirīgu mantisko tiesību sadalījumu, piemēram, datu bāzes www.dzivesstasts.lv gadījumā.

Fonogrammu producentam pieder tikai mantiskās tiesības uz fonogrammu vai tās kopiju, turklāt tās atšķiras no izpildītāja mantisko tiesību apjoma. Producentam ir tiesības darbu:

- izplatīt;
- padarīt pieejamu sabiedrībai pa vadiem vai citādā veidā tādējādi, ka tām var piekļūt individuāli izraudzītā vietā un individuāli izraudzītā laikā;
- nomāt, īrēt un publiski patapināt arī gadījumos, kad tās izplatījis pats fonogrammu producents vai tas noticis ar viņa piekrišanu;
- tieši vai netieši, īslaicīgi vai pastāvīgi reproducēt.

Fonogrammu producents līdzīgi kā izpildītājs vai darba autors var vienoties par citādu mantisko tiesību sadali.

Autortiesību likuma 53. pants nosaka raidorganizāciju tiesības uz to fiksētajiem darbiem:

- publiskot raidījumus par maksu vai vietās, kuras sabiedrībai pieejamas par maksu, vai vietās, kuru īpašnieki vai valdītāji izmanto raidījumus apmeklētāju piesaistīšanai;
- pārraidīt programmu nesošus signālus ar jebkuras citas raidorganizācijas, kabeļoperatora vai cita izplatītāja palīdzību;
- iegūt jebkādu fotogrāfisku raidījuma kadru attēlu (kadra fotogrāfiju), ja iegūšana netiek veikta personiskām vajadzībām, un pavairot vai izplatīt jebkuru šādu fotogrāfiju;
- retranslēt raidījumu pa kabeļiem;
- padarīt raidījumu vai tā fiksāciju pieejamu sabiedrībai pa vadiem vai citādā veidā tādējādi, ka tai var piekļūt individuāli izraudzītā vietā un individuāli izraudzītā laikā;
- fiksēt jebkuru raidījumu ar skaņas vai video aparatūras palīdzību, tieši vai netieši, īslaicīgi vai pastāvīgi reproducēt raidījuma fiksāciju un izplatīt jebkuru šādu fiksāciju.

Fonogrammas producentam ir ekskluzīvas tiesības atļaut vai aizliegt reproducēšanu, kas nozīmē fonogrammu pārrakstīšanu vai kopēšanu, piemēram, no kasetes uz kaseti, no CD uz CD, no interneta uz CD, no CD uz internetu utt. Fiziskām personām ir tiesības

veikt reproducēšanu no likumīgi iegūtas fonogrammas, neprasot atļauju fonogrammas producentam, bet, maksājot tam atlīdzību, izplatīšanu, kas nozīmē fonogrammas pārdošanu, padarīt pieejamu fonogrammu internetā, kas nozīmē fonogrammas ievietošanu internetā tādējādi, ka tā katra persona var tai piekļūt individuāli izraudzītā vietā, piemēram, mājās vai interneta kafejnīcā, un individuāli izraudzītā laikā[103].

6.7.4. *Secinājumi*

Lai ievērotu autortiesību subjektu personiskās un mantiskās tiesības ir nepieciešams noteikt, kam pieder autortiesības vai attiecīgi blakustiesības uz konkrēto darbu. To iespējams īstenot, ja darbā, kas jau iepriekš ir publicēts, ir norādīts autortiesību īpašnieks. Ja nav norādīts autors, bet kāds cits autortiesību pārņēmējs, tad jābūt piesardzīgiem un jāpārbauda autortiesību apjoms, kas pieder attiecīgajai personai. Ja autors nodevis mantiskās tiesības citai personai, tad nepieciešama abu autortiesību subjektu piekrišana, lai varētu darbu izmantot, izņemot gadījumu, kad līgums starp autoru un mantisko tiesību īpašnieku paredz iespēju darbu noteiktā formātā, apjomā un teritorijā izmantot, neprasot īpašu atļauju autoram.

Ja darbā nav norādes par to, kurš ir autortiesību īpašnieks, tad ir jāpārbauda, vai šī darba intereses pārstāv mantisko tiesību kolektīvā pārvaldījuma organizācijas.

Ja nav iespējams atrast autortiesību īpašnieku, tad attiecīgā darba izmantošana ir riskanta, jo nav iespējams iegūt atļauju izmantot darbu. Šis apgalvojums galvenokārt attiecas uz komerciālā nolūkā izmantojamiem darbiem, turpretī izglītības un zinātnes mērķiem darba autortiesību īpašnieka atļauja nav obligāta.

6.8. Juridiskie nosacījumi darbu ievietošanai Latviešu valodas korpusā

Pēc vienošanās par Latviešu valodas korpusā iekļaujamajiem darbiem jāorganizē darbs ar autortiesību subjektiem par darba izmantošanas atļauju. To iespējams realizēt, slēdzot attiecīgus līgumus ar autortiesību subjektiem vai saņemot licences.

Bernes konvencija nodrošina autortiesību aizsardzības minimālo termiņu – autora dzīves laikā un 50 gadus pēc autora nāves, tomēr valstu valdības drīkst palielināt šo termiņu nacionālajos autortiesību likumos, to pielāgojot vietējiem nosacījumiem. Aizsardzības ilgums mainās arī atkarībā no autora darba veida, darba radīšanas datuma un aplūkojamā tiesību veida. Eiropas Savienības Autortiesību direktīva, kas ievieš WIPO Autortiesību līgumu, ir pagarinājusi autoru tiesību aizsardzību Eiropas Savienībā līdz 70 gadiem pēc autora nāves. ASV *Sonny Bono* autortiesību termiņa pagarināšanas likums 1998. gadā pagarināja autortiesību aizsardzību līdz 70 gadiem pēc autora nāves, un no 75 līdz 95 gadiem darbiem kolektīviem darbiem[104].

[103] http://www.laipa.org – skatīts 16.06.2005.
[104] http://www.wipo.org – skatīts 08.05.2005.

Personiskajām tiesībām nav vienota termiņa: dažās valstīs (Francija), personiskās tiesības ir mūžīgas, turpretī citās valstīs personisko tiesību termiņš beidzas līdz ar mantiskajām tiesībām[105].

Darbu, kuru izmantošana bija aizliegta vai ierobežota no 1940. gada jūnija līdz 1990. gada maijam, publiskošana līdzīgi kā iepriekšējiem ir iespējama tikai ar autora vai to mantinieku atļauju. Svarīgi norādīt, ka šo autortiesību aizsardzības termiņa pagarinājumu nosaka tiesa. Kā piemērus var minēt šādus autorus, kuru autortiesību aizsardzības termiņš ir pagarināts ar tiesas spriedumu: E. Virza, A. Grīns, L. Breikšs.

Nozīmīgs ir arī jautājums par darba publiskošanu. Darba publiskošanas atļauju ir iespējams iegūt, slēdzot licences līgumu vai saņemot licenci no darba autora. Šī atļauja ir obligāti nepieciešama, ja Latviešu valodas korpusu plānots izmantot komerciāliem mērķiem. Arī gadījumos, kad likums paredz iespēju darbu izmantot bez autora piekrišanas un atlīdzības, piemēram, izglītības, pētniecības, informatīviem mērķiem, Latviešu valodas korpusam būtu vēlams slēgt vienošanos ar darba autoru par darba ievietošanu Latviešu valodas korpusā, jo:

- tas atvieglotu pušu tiesību un pienākumu sadali;
- tas būtu pierādījums darba izmantošanas atļaujai un darba izmantošanas mērķa noteikšanai.

Darbu bez autora piekrišanas un atlīdzības drīkst izmantot tikai tad, ja darbs iepriekš ir publiskots. Ja darbs ir tikai izziņots, tad viennozīmīgi ir nepieciešama autora atļauja, lai darbu varētu publiskot.

6.8.1. *Latvijā izziņotie darbi*

Latvijā izziņotie darbi pakļauti Latvijas Autortiesību likuma regulējumam. Tabulā parādīti galvenie nosacījumi, kas jāievēro, lai darbs tiktu izmantots likumīgi.

Autortiesību subjekti:	**Darbu izmantošanas iespējama, ja ir:**	**Nav nepieciešama atļauja vai samaksa:**
Darba autors	Licences līgums vai licence	pēc darba autora nāves pagājuši 70 gadi
Līdzautori	Licences līgums ar visu līdzautoru atļauju vai licence	pie nosacījuma, ka pēc pēdējā dzīvā līdzautora nāves pagājuši 70 gadi
Darbu, kuru izmantošana bija aizliegta vai ierobežota no 1940. gada jūnija līdz 1990. gada maijam, autori vai to mantinieki	Licences līgums vai licence	70 gadu termiņam pēc darba autora nāves klāt jārēķina aizlieguma vai ierobežojuma periods

[105] http://www.ipr-helpdesk.org/controlador.jsp?cuerpo=principal&seccion=guias&guia=guia2&len=en&redIzq&borde=no – skatīts 06.06.2005.

Anonīmo darbu autori	Licences līgums vai licence ar anonīmā darba autora pārstāvi, ja tāds ir	70 gadi no darba izziņošanas brīža
Atvasināto darbu autori	Licences līgums vai licence	Jebkurš no pirmajiem četriem uzskaitītajiem termiņiem atkarībā no darbu vai autoru veida
Autoru mantinieki	Licences līgums vai licence	Jebkurš no pirmajiem četriem uzskaitītajiem termiņiem atkarībā no darbu vai autoru veida
Mantisko tiesību kolektīvie pārvaldītāji	Licences līgums vai licence	Jebkurš no pirmajiem četriem uzskaitītajiem termiņiem atkarībā no darbu vai autoru veida
Izdevēji	Līgums ar izdevniecību, ja to paredz izdevniecības līgums, kas slēgts ar darba autoru. Licences līgums vai licence, slēdzami ar autoru, ja to neaizliedz izdevniecības līgums	Jebkurš no pirmajiem četriem uzskaitītajiem termiņiem atkarībā no darbu vai autoru veida
Darba devēji	Licences līgums vai licence, slēdzami ar autoru, ja to neaizliedz izdevniecības līgums, kas slēgts starp autoru un darba devēju	Jebkurš no pirmajiem četriem uzskaitītajiem termiņiem atkarībā no darbu vai autoru veida

Runas korpusa izveidei nepieciešams apkopot dažādu veidu darbus, kas pakļauti autortiesību aizsardzībai kā blakus tiesību objekti.

Blakustiesību subjekti:	**Darbu izmantošanas iespējama, ja ir:**	**Nav nepieciešama atļauja vai samaksa:**
Izpildītājs	Licences līgums vai licence, ja līgums starp darba radīšanā iesaistītajām personām neparedz citādi	50 gadi pēc pirmā izpildījuma vai arī 50 gadi pēc pirmās publicēšanas vai publiskošanas
Fonogrammu producents	Licences līgums vai licence, ja līgums starp darba radīšanā iesaistītajām personām neparedz citādi	50 gadi pēc tam, kad izdarīta fiksācija vai arī 50 gadi pēc pirmās publicēšanas vai publiskošanas
Raidorganizācijas	Licences līgums vai licence	50 gadi pēc raidījuma pirmās pārraides

Izpildītāju mantinieki	Licences līgums vai licence, ja līgums starp darba radīšanā iesaistītajām personām neparedz citādi	50 gadi pēc pirmā izpildījuma vai arī 50 gadi pēc pirmās publicēšanas vai publiskošanas
Fonogrammu producentu mantinieki	Licences līgums vai licence, ja līgums starp darba radīšanā iesaistītajām personām neparedz citādi	50 gadi pēc tam, kad izdarīta fiksācija vai arī 50 gadi pēc pirmās publicēšanas vai publiskošanas

Līdzautori. Mazliet atšķirīgas prasības ir attiecībā uz līdzautoru darbiem. Līdzautorus iespējams iedalīt divās grupās. Pirmie ir tādi līdzautori, kuru ieguldījumu darba radīšanā nav iespējams nodalīt. Tādā gadījumā ir nepieciešams slēgt līgumu vai saņemt licenci ar abu autoru piekrišanu. Otrā līdzautoru grupa ir autori, kuru radītais darbs neatkarīgi no otra autora radītā darba ir patstāvīgs un nodalāms ieguldījums kopējā darbā. No tā izriet, ka katrs autors atsevišķi pārstāv sava patstāvīgā darba daļu un līgumi vai licences jākārto ar katru autoru atsevišķi.

Anonīmie autori. Anonīmo darbu autoru atļauju parasti nav iespējams saņemt, taču anonīmā autora intereses var pārstāvēt, piemēram, darba izdevējs. Ja anonīmais autors atklāj sevi, tad autortiesību aizsardzības termiņš tiek rēķināts kā jebkuram citam darbu autoram.

Atvasināto darbu autori. Atvasināto darbu autori rada savus darbus, izmantojot kādu oriģināldarbu, taču šo autoru tiesības attiecas tikai uz viņu radīto darbu, bet ne uz izmantoto oriģināldarbu. Taču tas nenozīmē, ka atvasināto darbu nedrīkst izmantot bez oriģināldarbu autoru atļaujas. Līdzīgi kā atvasinātā darba autoram nav tiesību uz oriģināldarbu, tā oriģināldarba autoram nav tiesību uz atvasināto darbu tik tālu, cik atvasinātais darbs neaizskar oriģināldarba autora tiesības. Šīs oriģināldarba un atvasinātā darba autoru attiecības ir svarīgas gadījumā, ja tiek veidots Latviešu valodas paralēlais korpuss. Ar katra darba autoru jāslēdz atsevišķs līgums par konkrētā darba izmantošanu.

Mantinieki. Kā norādīts Autortiesību likuma 16. panta 1. daļā, autora mantiniekiem pāriet tiesības izziņot un izmantot darbu un saņemt atlīdzību par atļauju izmantot darbu, kā arī par darba izmantošanu. Autora mantiniekiem ir tiesības aizsargāt autora personiskās tiesības. Tas nozīmē, ka mantinieki, salīdzinot ar citiem autortiesību pārņēmējiem, iegūst lielāku rīcības brīvību attiecībā uz autora darba izmantošanu. Savukārt Autortiesību likuma 36. panta 2. daļa nosaka kārtību, kādā tiek iegūtas tiesības uz autora darbiem. Ja autors ir rakstījis testamentu, tad autors testamentā var norādīt personu, kurai viņš uztic autora tiesību aizsardzību pēc nāves. Ja nepastāv testaments vai testamentā nav norādīta attiecīgā persona, tad autora tiesības pēc viņa nāves realizē

viņa mantinieki (testamentārie, likumiskie). Svarīgi norādīt, ka mantojumam ir jābūt apstiprinātam, tikai pēc attiecīgā tiesas sprieduma mantinieki drīkst realizēt autora tiesību pārstāvību. Turklāt personas, kas mantojušas šīs tiesības, veic savas pilnvaras līdz mūža beigām. Līdz brīdim, kamēr mantojums nav vēl apstiprināts, tā pārvaldību veic ieceltais testamenta izpildītājs. Autoru mantinieks iegūtās tiesības drīkst nodot mantojumā nākamajām paaudzēm. Likums paredz arī gadījumu, kad autoram nav mantinieku un autortiesību aizsardzības termiņš vēl nav beidzies: „Ja mantinieku nav vai viņiem piederošo autortiesību termiņš izbeidzies, šo tiesību aizsardzību realizē autoru mantisko tiesību kolektīvā pārvaldījuma organizācija."

Mantinieku loku un mantošanas kārtību nosaka Latvijas Republikas Civillikums. Juridisks dokuments, kas apliecina to, ka konkrētā persona ir mantinieks, var būt:

- apliecība par tiesībām uz mantojumu, ko izdevis notariāta kantoris (līdz Civillikuma mantojuma tiesību daļas spēkā stāšanās brīdim – 1992. gada 1. septembrim),
- testaments, kuru tiesa pasludinājusi par spēkā stājušos,
- tiesas spriedums par mantošanas tiesību apstiprināšanu,
- mantošanas līgums.

Ja autoram ir vairāki mantinieki, viņi savas tiesības realizē kopīgi saskaņā ar Civillikuma 715. pantu. Likums nosaka, „ja mantojums piekritis vairākām personām kopīgi, tad viņas var vai nu valdīt to nedalīti vai prasīt tā dalīšanu" [Paklone, Lielkalns, Sosnovska, Tola 1997:137]. Visi mantinieki, kas manto tiesības uz autora darbu, šīs tiesības izmanto kopīgi, jo tiesību apjoma sadale nav iespējama. Turklāt gadījumos, kad autortiesības realizē vairāki mantinieki, ir nepieciešama visu mantinieku atļauja darbu izmantot. Gadījumā, ja kāds no mantiniekiem nepiekrīt atļaujas nosacījumiem, tad darba izmantošana nav iespējama.

Mantisko tiesību kolektīvie pārvaldītāji. Attiecības ar mantisko tiesību kolektīvo pārvaldītāju organizācijām aplūkotas **koncepcijas 6. 11. punktā.**

Izdevēji. Slēdzot līgumus ar izdevējiem par darbiem, ko paredzēts izmantot valodas korpusā, būtu jāievēro šādi nosacījumi, kas svarīgi, lai darbi tiktu izmatoti, ievērojot likumu:

- jāpārliecinās, vai izdevniecība ir noslēgusi izdevniecības līgumu ar raksta autoru,
- vai ievērota līguma forma,
- kāds ir izdevējam piešķirto tiesību apjoms.

Lielākajā daļā valstu ir noteikts, ka līgums starp autoru un izdevēju ir slēdzams, obligāti ievērojot rakstisku formu. Līguma rakstveida forma ir pierādījums noslēgtās vienošanās saturam. Tomēr rakstveida formas neievērošana ne vienmēr padara darījumu

par spēkā neesošu (*ad validatem*), bet ir domāta vienīgi pierādīšanas atvieglošanai (*ad probatem*) [106].

Autors, slēdzot izdevniecības līgumu, var nodot izdevējam autora mantiskās tiesības. Tas nozīmē, ka līgumu par darbu publiskošanu var slēgt ar izdevēju un nav nepieciešama autora atļauja. Taču pastāv vēl viens priekšnoteikums, lai autora radīto darbu varētu publiskot, izmantojot elektroniskos resursus (internets, CD formāts u. tml.), līgumā starp autoru un izdevēju jābūt norādītam atļautajam darba publiskošanas veidam, laikam un apjomam (teritorijai, kurā ir atļauts darbu publiskot). Pilnīgi pietiktu ar atrunu līgumā par to, ka izdevējs ir tiesīgs viņam nodotās tiesības uz darbu izmantošanu nodot tālāk.

Neprecīzi vai rakstveida formā nenoslēgti līgumi var būtiski ietekmēt arī darbu tālāku komerciālu izmantošanu. Tā, piemēram, ja autors cedējis izdevējam tiesības izdot laikrakstā viņa rakstu un tālāka (atkārtota) publicētā darba izmantošana nav precizēta, izdevējs nebūs tiesīgs laikraksta numuru publicēt internetā (it sevišķi, ja piekļūšana tam tiek piedāvāta kā maksas pakalpojums) vai izdot laikraksta numuru apkopojumu kompaktdiska formātā[107].

Ja izdevējam ir nodotas autora mantiskās tiesības, tad izdevējs ir tiesīgs sniegt prasības par autortiesību pārkāpumiem. Ja izdevniecības līgums neparedz mantisko tiesību pāreju, tad uz izdevēju var attiecināt tikai normas, kas saistītas ar konkurences pārkāpumiem. Tas nozīme, ka, slēdzot līgumu ar autoru par darba publiskošanu, būtu jāpārliecinās, vai netiek aizskartas izdevēja intereses. Šāda situācija ir iespējama, ja ir noslēgts izdevniecības līgums, kas paredz darba publiskošanas saskaņošanu ar izdevēju vai pat to aizliedz. Parasti gan izdevniecības līgumi nesatur šādas atrunas, bet, lai izvairītos no autortiesību pārkāpuma, ir nepieciešams vai nu līguma par darba publiskošanu, vai atsevišķā dokumentā nostiprināt autora vai izdevniecības apliecinājumu, ka netiek aizskartas attiecīgi izdevniecības vai autora tiesības uz darbu. Pastāv viedoklis, ka izvairīties no atbildības nepalīdzēs arī noslēgtie līgumi ar sākotnējo izdevēju, jo tie dod tikai tiesību uz regresa prasību par nodarītajiem zaudējumiem, nevis juridisku iespēju izvairīties no atbildības par autortiesību pārkāpumiem[108].

Darba devēji. Latvijā pastāv pretēji viedokļi par to, vai darba līgumā vai uzņēmuma līgumā var tikt iekļautas izdevniecības līgumā prasītās sastāvdaļas, tādējādi apejot nepieciešamību slēgt vairākus līgumus. Kā norāda M. Grudulis, darba līgums nekādā gadījumā nevar aizstāt izdevniecības līgumu (Autortiesību likuma 12. pants), un tādējādi līgums, ar kuru tiek pieņemts darbā štata žurnālists, nevar aizstāt izdevniecības līgumu, kurš jāslēdz par katru darbu un tā izmantošanu atsevišķi [109]. Savukārt

[106] http://www.lpia.lv/?id=373 – skatīts 05.07.2005.
[107] http://www.lpia.lv/?id=373 – skatīts 05.07.2005.
[108] Ibid.
[109] Ibid.

J. Rozenfelds uzskata, ka no normas burtiskā teksta it kā izriet secinājums, ka līdztekus darba līgumam paredzēta vēl kāda cita īpaša vienošanās par darba līguma ietvaros radītā darba nodošanu darba devējam. Jādomā tomēr, ka tāds nav bijis likumdevēja nolūks, drīzāk jāpieņem, ka šādos gadījumos, kad jaunrades akts ietilpst autora darba pienākumos, mantiskās tiesības uz autora radīto darbu pāriet darba devējam, tieši pamatojoties uz to pašu darba līgumu [Rozenfelds 2004:47].

Lai izvairītos no autortiesību pārkāpuma, šādos gadījumos ir ieteicams pārliecināties par to, vai darba devējam noslēgtajos līgumos vai tas būtu darba līgums, vai izdevniecības līgums ir atrunāta autora mantisko tiesību pāreja. Ja šādas atrunas līgumā nav, tad, ievērojot Autortiesību likuma mērķi aizsargāt autoru tiesības, visas neskaidrības jātulko par labu autoram. Attiecībā uz uzņēmuma līgumu – tas pēc būtības atbilst Autortiesību likuma 13. panta normai „Autora līgums par pasūtītu darbu", līdz ar to nepastāv ierobežojums slēgt uzņēmuma līgumu par darba radīšanu.

Blakustiesību subjekti. Līdzīgi kā ar pārējiem autortiesību aizsargātajiem darbiem, lai izmantotu jebkuru blakustiesību aizsargātu darbu, ir nepieciešama darba radītāju atļauja. Tā kā šāds darbs pēc savas būtības ir salikts darbs, jo darba radīšanas procesā ir piedalījušās vairākas personas, tad iespējams, ka atļauja jāsniedz katram darba tapšanā iesaistītajam. Līdzīgi kā autortiesību aizsargātajiem darbiem arī blakustiesību aizsargātajiem darbiem ir svarīgi noskaidrot izmantošanas mērķi, lai varētu noslēgt attiecīgu licences līgumu vai saņemt licenci darba izmantošanai. Kā norādīts iepriekš, ļoti liela daļa ierakstu latviešu valodā ir veidoti tieši izglītības un pētniecības mērķiem, tāpēc šādu darbu izmantošana komerciālos nolūkos ir neiespējama.

Runas korpusa ietvertie darbi skar galvenokārt šādu blakustiesību subjektu intereses:

- izpildītājs,
- fonogrammu producents,
- raidorganizācijas.

Izpildītājam tāpat kā fonogrammu producentam pieder mantiskās tiesības uz radīto darbu, taču pirms darba radīšanas iespējams slēgt līgumu par mantisko tiesību pāreju vienai vai otrai iesaistītajai pusei. Savukārt personiskās tiesības ir neatņemamas un tiek saglabātas atbilstoši likumā noteiktajam apjomam darba autoram un izpildītājam. Izpildītājs var nodot fonogrammu producentam tikai daļu savu tiesību. Gadījumā, ja izpildītājs savas iznomātāja tiesības attiecībā uz fonogrammas oriģinālu vai kopiju ir nodevis fonogrammas producentam, izpildītājs patur tiesības saņemt taisnīgu atlīdzību par nomu. Vienošanās par izpildītāja atteikšanos no tiesībām saņemt atlīdzību uz turpmāko laiku nav spēkā. Izpildītājam pieder tiesības atļaut vai aizliegt izpildījuma fiksāciju. Bez izpildītāja atļaujas nevienam nav tiesības fiksēt izpildījumu. Gadījumos, kad izpildījums jau ir fiksēts (saņemot attiecīgu izpildītāja atļauju), izpildītājam rodas tiesības atļaut vai aizliegt izpildījuma pārdošanu, padarīt to pieejamu inetrnetā.

Parasti valstīs, kurās izpildītājiem nav savas arodbiedrības vai interešu organizācijas, šīs tiesības administrē paši izpildītāji, slēdzot līgumus ar fonogrammas producentu un vienojoties par izpildījuma fiksāciju izmantošanas nosacījumiem norādītajos veidos.

Izpildītājam ir tiesības saņemt atlīdzību par fiksētā izpildījuma:

- raidīšanu;
- retranslāciju pa kabeļiem;
- atskaņošanu publiskās vietās;
- reproducēšanu personiskām vajadzībām (nesēja atlīdzība).

Tas nozīmē, ka jebkurš drīkst izmantot izpildījuma fiksāciju, neprasot atļauju izpildītājam, taču viņam ir jāsamaksā atlīdzība par tās izmantošanu. Saskaņā ar Civillikumu saistības maksāt atlīdzību rodas brīdī, kad izmantotājs izpildījuma fiksāciju atskaņo. Šie paši nosacījumi attiecas ne tikai uz Latvijas, bet arī uz ārvalstu izpildītājiem. Parasti šīs tiesības izpildītāji neadministrē paši, bet uztic pašu izveidotai mantisko tiesību kolektīvā pārvaldījuma organizācijai[110].

Producents ir fiziskā vai juridiskā persona, kas organizē un finansē audiovizuālu darbu radīšanu (filmas producents) vai fiziskā vai juridiskā persona, kas veic autora darba izpildījuma skaņu vai skaņu atveidojuma pirmo fiksāciju un ir atbildīga par tās pabeigšanu (fonogrammas producents). Tieši producents ir atbildīgs par autortiesību ievērošanu šajos darbos. Viņš ir tas, kurš no autora vai to pārstāvošās autortiesību organizācijas prasa un saņem atļauju darbu izmantošanai, kā arī samaksā autoratlīdzību[111]. Fonogrammu producentam tiesības uz darbu rodas pēc līguma noslēgšanas ar autoru un izpildītāju, noslēgtais līgums nosaka to tiesību apjomu, kura ietvaros fonogrammu producents ir tiesīgs pārstāvēt autortiesības un blakustiesības. Lai darbu varētu izmantot, fonogrammu producentam ir pienākums saņemt atļauju ne tikai no darba autora, bet arī no izpildītāja. Atļaujas nav savstarpēji aizvietojamas, tas nozīmē, ka nepietiek tikai ar vienu atļauju.

Personai, kura vēlas izmantot fonogrammas, reproducējot, izplatot, ievietojot tās internetā, vai citā veidā, ir jāsaņem atļauja šādu darbību veikšanai no fonogrammu producenta. Bez šādas atļaujas (licences) saņemšanas fonogrammas izmantot ir aizliegts un jebkura izmantošana, ja nav saņemta atlīdzība, ir uzskatāma par fonogrammu producenta tiesību pārkāpumu, par ko var saņemt sodu likumos noteiktā kārtībā un/vai maksāt kompensāciju vai atlīdzināt zaudējumus fonogrammas producentam[112].

[110] http://www.laipa.org – skatīts 17.05.2005.
[111] http://www.lpia.lv/?id=373 – skatīts 05.07.2005.
[112] http://www.laipa.org – skatīts 17.05.2005.

Raidorganizāciju veidotie raidījumi un jebkuri skaņu ieraksti ir to īpašums, tāpēc šo darbu izmantošanai ir nepieciešama attiecīgas raidorganizācijas atļauja. Arī raidorganizāciju radītajos darbos ir iesaistītas gan darbu autoru, gan izpildītāju tiesības un intereses.

Līdzīgi kā autortiesību aizsardzības gadījumā arī blakustiesību aizsardzības gadījumā pastāv izpildītāju, fonogrammu producentu un raidorganizāciju tiesību ierobežojumi likumā noteiktos gadījumos. Bez blakustiesību subjektu piekrišanas un bez atlīdzības samaksas pieļaujams izmantot blakustiesību objektu, kā arī to fiksēt šādos gadījumos:

1) personiskām vajadzībām;
2) nelielos fragmentos, kas iekļauti ziņu raidījumos un aktuālo notikumu apskatos;
3) izglītības un pētniecības nolūkos;
4) citos nolūkos, kas attiecībā uz darbu autoru mantisko tiesību ierobežošanu noteikti likumā.

Izpildītāju un fonogrammu producentu mantinieki manto blakus tiesības tādā pašā kārtībā kā cita autortiesību īpašuma mantošanas gadījumā. Likums nenorāda mantojamo tiesību apjomu, līdz ar to nav pamats domāt, ka blakustiesību mantošana ir ierobežota ar kādu nosacījumu. Likums norāda, ka izpildītājam pieder gan personiskās tiesības, gan mantiskās tiesības, no likuma būtības un mērķa izriet, ka personiskās tiesības arī blakustiesību subjektiem nav nododamas citām personām, tās pieder tikai izpildītājam. Pēc līdzības ar autora mantinieku tiesībām arī izpildītāju mantiniekiem ir tiesības aizsargāt izpildītāja personiskās tiesības.

Lai blakustiesību subjekti būtu aizsargāti ar Latvijas Republikas likumiem pašiem subjektiem vai to radītajiem darbiem ir jābūt saistītiem ar Latviju. Šo blakustiesību subjektu piesaistes nosacījumi atrodami Autortiesību likuma 56. pantā. Attiecībā uz raidorganizācijām ir prasība, ka oficiālajai atrašanās vietai ir jābūt Latvijā.

6.8.2. *Ārpus Latvijas Republikas izziņotiem darbiem*

Lai gan ir pieņemtas starptautiskas konvencijas ar mērķi harmonizēt autortiesības visā pasaulē, tomēr vēl joprojām pastāv atšķirīgas autortiesību aizsardzības sistēmas.

Eiropas sistēma paredz, ka autortiesības pastāv bez īpašas reģistrācijas, taču ASV tomēr vēl pastāv reģistrācijas prasība, lai gan tas vairs netiek uzskatīts par priekšnosacījumu, lai darbs tiktu uzskatīts par autortiesību objektu un būtu aizsargāts. Reģistrācija atvieglo darba autora tiesību aizsardzību.

ASV intelektuālais īpašums tiek aizsargāts ar t. s *copyright* tiesībām, kas aizsargā ne tikai autorus, bet arī personas, kuras veikušas ieguldījumu darba radīšanā. Tas nozīmē,

ka tiek aizsargāts plašāks darbu loks, jo tiek aizsargāti arī tie darbi, kas nav jaunrades rezultāts, bet to sagatavošanā ir veikts liels ieguldījums. ASV sistēma aizsargā arī plašāku personu loku, kas iesaistīti darba veidošanā, jo nosaka, ka tiek aizsargāta jebkura persona, kas veikusi ieguldījumu darba tapšanā (gan autors, gan izdevējs, gan producents u. c.). ASV ir noteiktas stingras prasības, kas jāievēro, slēdzot izdevniecības līgumu. Līgums sastāv no daudzām būtiskām sastāvdaļām. Šie līgumi jāslēdz rakstveidā.

Pretēja situācija ir Eiropas ietvaros, kur tiek vērtēta darba jaunrade vai oriģinalitāte, proti, tiek izvirzīti nosacījumi darbam, lai to varētu uzskatīt par autortiesību objektu. Darba atbilstības noteikšanu strīdu gadījumos veic tiesa. Atkarībā no valsts, kurā tiek skatīts strīds, vērojama atšķirīga izvērtēšana. Savdabīgākās formas vērojamas Francijas un Anglijas tiesu praksē. Taču kopīga pazīme – Eiropas sistēma uzsver, ka jārūpējas par autoru tiesībām, kas sastāv gan no personiskajām (nemantiskajām), gan no mantiskajām tiesībām.

Viena no kritērija radikālākajām formām – subjektīvā oriģinalitāte (Francija) – nosaka, ka darbam jāatspoguļo autora personalitāte (l'empreinte de la personalitē). Darba oriģinalitāti vai neoriģinalitāti iespējams konstatēt tikai katrā konkrētā gadījumā, un to parasti veic tiesa, izskatot konkrētu gadījumu[113].

Otra radikālā oriģinalitātes forma – objektīvā oriģinalitāte (Anglija) – tiek interpretēta kā darba novitāte, un līdz ar to darbam, lai tas tiktu aizsargāts, jābūt vienīgi jaunam, tas ir, tādam, kas neatkārto cita autora agrāk radītu darbu[114].

Atšķirīgu valstu likumos līdzīgi kā Latvijas Autortiesību likumā ir paredzētas izņēmuma tiesības, kad darbu var izmantot pie atvieglotiem nosacījumiem, piemēram, ir valstis, kas šādus izņēmumus piemēro attiecībā uz parodijām, citātiem. 2001. gada 22. maija Direktīva 2001/29/EK par dažu autortiesību un blakustiesību aspektu saskaņošanu informācijas sabiedrībā daļēji mēģina apkopot un atrast kopīgos, nacionālajos likumos paredzētos izņēmumus.

Atsevišķās valstīs, piemēram, Nīderlandē un Lielbritānijā, darba devējs tiek uzskatīts par vienīgo autortiesību īpašnieku, ja darbu radījis viņa darbinieks, turklāt nav nepieciešams speciāls līgums, kas norāda uz autora tiesību piederību.

Līdzautoru darbi (*Collective works, l'oeuvre collective*) dažās valstīs (Francija, Itālija, Luksemburga, Portugāle, Spānija) gadījumos, kad katra autora ieguldījumu koordinē cita persona (gan fiziska, gan juridiska persona), nozīmē, to, ka autortiesības uz darbu piederēs tieši koordinatoram tik tālu, cik katra autora ieguldījumu nav iespējams personificēt, t. i., nodalīt no kopējā darba.

[113] http://www.lpia.lv/?id=373 – skatīts 05.07.2005.
[114] Ibid.

Darba tiesisko attiecību gadījumā arī pastāv atšķirīgs regulējums attiecībā uz darba laikā vai uzdevumā radītajiem darbiem. Šie darbi arī ir autortiesību objekts pat tad, ja darba devēja un darba ņēmēja autortiesību pāreju neregulē līgums.

Lielākajā daļā Eiropas Savienības valstu darba devējs ir autortiesību īpašnieks darbiem, ko radījuši darba ņēmēji, pildot savus darba pienākumus. Tomēr darba devēja tiesības izmantot darbu ir ierobežotas, jo izmantošana iespējama tikai attiecīgā darba uzdevuma veikšanai, lai nodrošinātu normālu biznesa darbību. Jebkurā citā darba ņēmēja radīta darba izmantošanas gadījumā vispirms ir nepieciešama darba radītāja atļauja. Tā kā personiskās tiesības nav atsavināmas, tad mantisko tiesību pāreja darba devēju neatbrīvo no atļaujas saņemšanas no darba autora, ja ir plānots cits darba izmantošanas veids vai ir nepieciešama darba modificēšana[115].

Dažās valstīs (galvenokārt Francijā, Itālijā, Beļģijā, Luksemburgā un Portugālē), autors paliek autortiesību īpašnieks, pat tad, ja darbs ir radīts darba līguma izpildes gaitā. Francijā un Luksemburgā tomēr pastāv speciāls režīms kolektīviem darbiem. Tas attiecas uz darbiem, ko radījušas vairākas personas, kuras bijušas pakļautas fiziskas vai juridiskas personas norādījumiem. Pie šādiem nosacījumiem attiecīgā fiziskā vai juridiskā persona tiek uzskatīta par sākotnējo autortiesību īpašnieku darbam, kas satur vairāku citu personu ieguldījumu. Šāds nosacījums ir izdevīgs darba devējam, kas darbu var organizēt tā, ka jebkurš darbs, kas radies darba izpildes procesā tiktu uzskatīts par darba devēja intelektuālo īpašumu[116].

Darba ņēmēja autortiesības arī ir nostiprinātas vairāku valstu likumos (piemēram, Latvijā). Tādā gadījumā autortiesību pāreja uz darbiem, kas radīti darba līguma izpildes procesā, ir iespējama, to atrunājot līgumā, kas noslēgts starp darba devēju un darba ņēmēju. Turklāt pastāv šādi nosacījumi: līgumam ir jābūt noslēgtam rakstiski, atrunu par autortiesību pāreju var atrunāt arī darba līgumā vai arī atsevišķā līguma, kas regulē darba ņēmēja radīto darbu izmantošanas kārtību.

Katras valsts autortiesību likumi nosaka formas prasības, kas saistītas ar autortiesību pāreju. Tā kā šie risinājumi Eiropas Savienībā nav harmonizēti, tad gadījumos, kad jāslēdz starptautiski privāttiesību līgumi par darbu izmantošanu, ir nepieciešams rūpīgi izpētīt citas valsts autortiesību regulējumu.

6.9. Līguma būtiskās sastāvdaļas

Kā norādīts jau iepriekš, par katru darba izmantošanas reizi un uz katru izmantošanas veidu ir jāsaņem autortiesību vai blakustiesību subjektu atļauja. Viena no atļaujas formām ir licences līgums (turpmāk saukts līgums), bet otra ir licence.

[115] http://www.ipr-helpdesk.org/controlador.jsp?cuerpo=principal&seccion=guias&guia=guia3&len=en – skatīts 06.06.2005.
[116] http://www.ipr-helpdesk.org/controlador.jsp?cuerpo=principal&seccion=guias&guia=guia3&len=en – skatīts 06.06.2005.

Atšķirībā no licences līgums parasti ietver plašāku pušu tiesību un pienākumu aprakstu. Līguma teksts var būt gan īsāks, gan garāks, taču līgums uzskatāms par galīgi noslēgtu, kad puses vienojušās un līgums satur visas līguma būtiskās sastāvdaļas. Kā nosaka Civillikuma 1470. pants „būtisks darījumā ir viss tas, kas nepieciešams tā jēdzienam un bez kā arī pats nodomātais darījums nav iespējams".

Autortiesību likuma 41. pants nosaka, ka licences līgums ir līgums, ar kuru viena puse – autortiesību subjekts – dod atļauju otrai pusei – darba izmantotājam – izmantot darbu un nosaka darba izmantošanas veidu, vienojoties par izmantošanas noteikumiem, atlīdzības lielumu, tās izmaksāšanas kārtību un termiņu.

Šis uzskaitījums arī faktiski jāuzskata par autora līguma būtiskajiem nosacījumiem, par kuriem pusēm noteikti jāvienojas, slēdzot līgumu [Paklone, Lielkalns, Sosnovska, Tola 1997:141]. Tas nozīmē, ka, slēdzot līgumu par darba izmantošanu, pusēm ir jāvienojas par līguma priekšmetu:

1) kādu darbu ar līgumu atļauj izmantot,
2) kāds ir atļautais darba izmantošanas veids,
3) kādi ir izmantošanas noteikumi.

Pusēm jāvienojas arī par samaksas kārtību:

1) kāds ir atlīdzības lielums,
2) atlīdzības izmaksāšanas kārtība,
3) samaksas termiņš.

Šis likuma formulējums attiecināms arī uz blakustiesību subjektu radītajiem darbiem. Līgumu var izmantot arī kā papildinājumu izdotajai licencei. Likums paredz, ka līgumā var atrunāt darbu izmantošanas veidus, izmantotāja tiesības nodot licenci trešajām personām. Gadījumā, ja līgums paredz iespēju nodot izsniegto licenci trešajām personām, tad līgums ir sniedzis tā saukto sublicenci, neskatoties uz to, ka licence neparedz šādu iespēju. Jebkuras no mantiskajām tiesībām autortiesību vai blakustiesību subjekti var nodot pilnībā vai daļēji darba izmantotājam. Līgumu var slēgt gan mutvārdos, gan rakstveidā. Lai gan paredzēta līguma slēgšanas brīvība attiecībā uz līguma formu, tomēr likums nosaka tos gadījumus, kad līgums obligāti slēdzams rakstveidā:

1) izdevniecības līgums;
2) līgums par darba publiskošanu;
3) līgums par audiovizuāla darba radīšanu;
4) līgums, kas nosaka tādas tiesības, kādas ietvertas vispārējā vai izņēmuma licencē.

Neskatoties uz līguma formas brīvību, tieši rakstiskā forma spēj nodrošināt pietiekamu un drošu pierādījumu pušu vienošanās esamībai. Mutvārdu formas izmantošana iespējama pušu savstarpējas uzticības gadījumā. Savukārt rakstiskā forma mazina iespēju, ka puses atšķirīgi saprot savus pienākumus un tiesības.

6.9.1. *Darba apjoms*

Lai uzskatītu, ka starp pusēm ir noslēgts līgums, tam jābūt pietiekami konkrētam. Viens no priekšnoteikumiem ir līguma priekšmeta esamība. Darbu izmantošanas līgumos līguma priekšmets ir autora vai blakustiesību subjektu darba gala rezultāts. Līgumā norādei uz darbu ir jābūt pietiekami konkrētai, proti, jānorāda darba apjoms, jo ar līgumu var atļaut izmantot vienu vai vairākus darbus, vai nu visu darbu vai konkrētu darba daļu. Taču līguma priekšmets nav pilnīgi konkrēts, ja nav norādes uz atļauto darba izmantošanas veidu. Likums gadījumos, kad nav norādes uz darba izmantošanas veidu, nosaka, „ja līgumā nav norādījumu par darba izmantošanas veidu, darba izmantotāja tiesības tiek ierobežotas ar tām darbībām, kas izriet no līguma un ir nepieciešamas līguma mērķa sasniegšanai". Šāds formulējums tomēr var radīt strīdu, gadījumos, kad līgumā nav noteikts mērķis. Ja norādīts darba izmantošanas veids, līgumslēdzējas puses var noteikt kārtību, kādā notiek darba izmantošana, piemēram, norādot, vai darbs tiks izmantots komerciālos nolūkos vai gluži pretēji darbs tiks izmantots nekomerciālos nolūkos izglītības un pētniecības mērķiem. Papildus autors vai blakustiesību subjekts var noteikt prasības, kas attiecas uz nodotā darba tālāku izmantošanu, neprasot tiešu atļauju attiecīgajam subjektam.

6.9.2. *Darbu teritoriālā pieejamība un autortiesības*

Līgumā ir svarīgi norādīt darba izmantošanas atļaujas darbības teritoriju īpaši interneta tīkla globālā rakstura dēļ. Autortiesību likuma 45. pants paredz, ka līgumā var norādīt teritoriju, kurā spēkā ir attiecīgā darba izmantošanas atļauja. Gadījumā, ja puses nav norādījušas teritoriju, tad tiek uzskatīts, ka atļauja ir spēkā tikai tajā valstī, kurā noslēgts līgums. Gadījumā, kad autors vai blakustiesību subjekts dod atļauju darba izmantošanai globalajā interneta tīkla, tad jau pēc izmantošanas veida var saprast, ka darbs būs pieejams jebkurā valstī, jebkurā laikā, taču līgumā tas būtu īpaši jāuzsver.

Personu loku, kam būs pieeja Latviešu valodas korpusam var ierobežot ar tehniskiem līdzekļiem:

- nosakot, ka jebkuram Latviešu valodas tekstu korpusa lietotājam ir nepieciešama lietotāja parole, ko var iegūt tikai un vienīgi slēdzot līgumu par Latviešu valodas korpusa izmantošanu,
- ļaujot korpusam pieslēgties tikai lietotājiem no Latvijas (pēc IP adresēm). Taču šajā gadījumā tiktu izslēgta iespēja, ka korpusam pieslēdzas Latvijas iedzīvotāji, kas atrodas ārvalstīs.

Prognozējams, ka problēmas var rasties ar atļauju dabūšanu no attiecīgajiem autortiesību un blakustiesību subjektiem, jo bieži vien tieši darbu elektroniskās versijas ir visvieglāk izmantot pretēji autoru un blakustiesību subjektu interesēm.

6.9.3. *Darba publiskošanas termiņš*

Līgumā var norādīt termiņu, cik ilgi darbojas autortiesību vai blakustiesību subjektu izsniegtā atļauja. Tomēr tas nav obligāts nosacījums. Pusēm ir izdevīgi nenorādīt konkrētu darba izmantošanas termiņu, jo tādā gadījumā pusēm nav jāseko līdzi noteiktā termiņa beigām. Ja puses līgumā nav noteikušas darba izmantošanas termiņu, tad jāvadās pēc likumā noteiktā, t. i., atļauja darbu izmantot darbojas tik ilgi, kamēr viena no pusēm līgumu neatsauc.

Lai atsauktu līgumu, ir nepieciešams ievērot likumā noteikto termiņu, proti, par līguma laušanu otra puse ir jābrīdina sešus mēnešus pirms līgumā noteiktās atļaujas atsaukšanas fakta. Šo kārtību nosaka Autortiesību likuma 44. panta 2. daļa. Tas pats pants nosaka, ka šī daļa ir imperatīva un tās atcelšana ar pušu vienošanos nav iespējama.

6.9.4. *Samaksas kārtība (noteikta maksa gadā, maksa par ikreizēju darba apskati internetā)*

Samaksa ir piemērojama, gadījumos:

1) ja tiek veikta komerciāla Latviešu valodas korpusa izmantošana un
2) ja uz korpusu nav attiecināmi Autortiesību likuma noteiktie izņēmumi (kad nav nepieciešama autora atļauja un nav jāmaksā autoratlīdzība).

Taču likums nepieļauj situāciju, kad Latviešu valodas korpuss tiek izveidots, lai realizētu Autortiesību likumā noteiktos izņēmumus, taču no lietotājiem prasītu samaksa par korpusa izmantošanu. Ja izņēmums tiek attiecināts uz darba izmantošanu informatīviem mērķiem, ko nosaka Autortiesību likuma 20. pants, tad pastāv iespēja saņemt atlīdzību par izveidoto darbu. Taču, ja darbs tiek izmantots izgītības un pētniecības mērķiem, kā to nosaka Autortiesību likuma 21. pants, tad nav pieļaujams komerciāls nolūks. Tomēr jāizvērtē termina „komerciāls nolūks" saturs. Neraugoties uz to, ka valodas korpuss tiks izveidots izglītības un pētniecības mērķiem, gan tā izstrādē, gan turpmākajā uzturēšanā ir jāiegulda ievērojami finanšu līdzekļi. Tāpēc būtu jāizvērtē iespēja segt vismaz korpusa uzturēšanas pašizmaksu no lietotāju līdzekļiem, ja attiecīgais valsts finansējums netiek nodrošināts. Šajā gadījumā nedrīkst būt runa par peļņas gūšanu (komerciāls nolūks), bet gan tikai par izdevumu segšanu.

Ja Latviešu valodas korpuss tiek veidots ar komerciālu mērķi, tad tekstu korpusa ietvaros ir svarīgi noskaidrot autoratlīdzības samaksas kārtību un apjomu, bet runas korpusa ietvaros blakustiesību subjektu atlīdzību. Būtiski norādīt, ka autoratlīdzības ir

attiecināmas uz darbu autoriem un neskar blakustiesību subjektu intereses. Blakustiesību subjektiem savukārt pienākas atlīdzība par viņu radītā darba ieguldījumu.

1886. gada 9. augusta Bernes konvencija par literatūras un mākslas darbu aizsardzību nosaka, ka darba autors ir pats tiesīgs izlemt samaksas apmēru, ko vēlas saņemt par sava darba izmantošanu. Mantisko tiesību kolektīvā pārvaldījuma organizācijas realizē darbu radītāju noteikto samaksu iekasēšanu. Pastāv iespēja, ka organizācija pati noteikusi atlīdzības lielumu, kas maksājams par darbu izmantošanu. To parasti piemēro gadījumos, ja organizācija ar darba radītāju ir vienojusies par šādu atlīdzības lielumu vai ja pēc darba radītāja nāves autortiesību vai blakustiesību mantisko tiesību pārvaldījumu pārņem šī organizācija.

Ja licences līgumā atlīdzības lielums nav konkretizēts, strīda gadījumā to nosaka tiesa pēc saviem ieskatiem.

Likums bieži vien nepieļauj noteikt vienreizēju atlīdzību autoram, bet paredz to noteikt proporcionāli izdoto eksemplāru mazumtirdzniecības cenai. Ir sastopama arī jauktās atlīdzības kārtība (piemēram, Francijā), kur vienreizēju atlīdzību atļauts izmaksāt, slēdzot līgumus par atsevišķu kategoriju darbu izmantošanu (piemēram, enciklopēdijas), atsevišķiem izmantošanas veidiem (reklāma), līgumos ar ārvalstniekiem vai situācijās, kad vienkārši neatmaksājas veikt aprēķinus proporcionālās atlīdzības iekasēšanai[117].

Jāņem arī vērā, ka starptautiskajos līgumos ietvertie principi ir pret legālo licenču principa izmantošanu autortiesībās. Tas nozīmē: izdevējs nevar attaisnoties ar to, ka ir gatavs izmaksāt autoram atlīdzību par viņa raksta izmantošanu atkārtoti, jo tikai autors ir tiesīgs lemt par sava darba ekonomisko izmantošanu arī pēc tā publicēšanas. Tādējādi, ja no kompaktdiskā publicētā laikraksta desmit gadu numuru apkopojuma viena raksta atkārtota publikācija šādā formātā nav attiecīgi juridiski noformēta, kā kontrafakta produkcija var tikt konfiscēta visa disku partija, kas neapšaubāmi līdzi nesīs ievērojamus zaudējumus[118].

Samaksas apmērs dažādiem autoriem un dažādiem darbiem var būt atšķirīgs. Piemēram, gadījumā, jā tiek lūgts izsniegt licenci AKKA/LAA, bet pastāv iespēja (blakustiesību subjektu aizsardzības gadījuma), ka tiek noteikts vienots tarifs, kas tiek attiecināts uz konkrētās valsts teritoriju.

Fonogrammas producentam ir tiesības saņemt atlīdzību par komerciālos nolūkos publicētas fonogrammas raidīšanu, tostarp raidīšanu internetā. Tas nozīmē, ka izmantotājs drīkst izmantot fonogrammas, neprasot atļauju fonogrammu producentam, taču tam ir jāsamaksā atlīdzība par fonogrammas izmantošanu. Saskaņā ar Civillikumu saistības maksāt atlīdzību rodas brīdī, kad izmantotājs fonogrammu izmanto. Šie paši

[117] http://www.lpia.lv/?id=373 – skatīts 05.07.2005.
[118] Ibid.

nosacījumi attiecas ne tikai uz Latvijas, bet arī uz ārvalstu fonogrammu producentiem. Parasti šīs tiesības fonogrammas producenti neadministrē paši, bet uztic pašu izveidotai mantisko tiesību kolektīvā pārvaldījuma organizācijai[119].

6.10. Licence (vienkāršā, izņēmuma, vispārējā)

Licence ir viena no atļaujas formām, kas darba izmantotājam ir jāsaņem attiecīgi no autortiesību vai blakustiesību darbu radītājiem. Lai varētu izmantot darbu, atļauja ir jāsaņem vai nu licences līguma, vai licences formā. Nepastāv obligāts nosacījums, ka jābūt gan licences līgumam, gan licencei, tomēr pastāv iespēja, ka puses vienojas, ka tiek izsniegta licence, bet papildus slēdz līgumu, lai vienotos par citiem ar darba izmantošanu saistītiem nosacījumiem. Licence obligāti ir izsniedzama rakstveidā. Licencē jānorāda katra darba izmantošanas reize, darba izmantošanas noteikumi un katru izmantošanas veidu, ko atļauj darba radītājs. Autortiesību likums nosaka, ka pastāv trīs licences veidi:

1) vienkāršā licence dod licences saņēmējam tiesības veikt tajā norādītās darbības vienlaikus ar autoru vai citām personām, kuras arī saņēmušas vai saņems attiecīgo licenci;
2) izņēmuma licence dod tiesības veikt tajā norādītās darbības vienīgi licences saņēmējam;
3) vispārējo licenci izsniedz autoru mantisko tiesību kolektīvā pārvaldījuma organizācija, un šī licence dod tiesības izmantot visu to autoru darbus, kurus pārstāv šī organizācija.

Licencē var norādīt laiku, cik ilgi darbojas autortiesību vai blakustiesību subjektu izsniegtā atļauja. Gadījumā, ja licencē nav termiņa atrunas, tiek uzskatīts, ka licence darbojas tik ilgi, kamēr viena no pusēm licenci neatsauc. Lai atsauktu licenci, ir nepieciešams ievērot likumā noteikto termiņu, proti, par licences atsaukšanu otra puse ir jābrīdina sešus mēnešus pirms licences atsaukšanas fakta. Šo kārtību nosaka Autortiesību likuma 44. panta 2. daļa, tas pats pants nosaka, ka šī daļa ir imperatīva un tās atcelšana ar pušu vienošanos nav iespējama.

Tāpat kā līguma slēgšanas gadījumā arī licences izsniegšanas gadījumā jāpievērš vērība atļaujas teritoriālajam spēkam. Autortiesību likuma 45. pants paredz, ka licencē var norādīt teritoriju, kurā spēkā ir attiecīgā darba izmantošanas atļauja, gadījumā, ja teritorija nav norādīta, tad tiek uzskatīts, ka atļauja ir spēkā tikai tajā valstī, kurā izsniegta licence.

[119] http://www.laipa.org – skatīts 12.05.2005.

6.11. Mantisko tiesību kolektīvie pārvaldītāji

Viens no autortiesību un blakustiesību subjektiem ir autoru izveidotas mantisko tiesību kolektīvā pārvaldījuma organizācijas. Tās izveidotas ar mērķi veikt kolektīvu autortiesību pārvaldīšanu. Jau pats nosaukums liecina, ka šīs organizācijas rūpējas par autortiesību mantisko tiesību atbilstošu izmantošanu un autoratlīdzību iekasēšanu. Līdzīgi darbojas blakustiesību kolektīvā pārvaldījuma organizācijas, tikai to izveidē nepiedalās raidsabiedrības.

Parasti autortiesību un blakustiesību pārvaldījumu organizācijas darbojas atsevišķi, tas skaidrojams ar autortiesību un blakustiesību aizsardzības atšķirīgajiem objektiem un subjektiem. Latviešu valodas korpusa izstrādes laikā iespējams nodalīt teksta un runas korpusa daļas. Teksta korpusa izstrādes gadījumā būs jāvēršas pie autortiesību kolektīvā pārvaldījuma organizācijas, bet runas korpusa izstrādes gadījumā pie blakustiesību kolektīvā pārvaldījuma organizācijas.

Mantisko tiesību kolektīvā pārvaldījuma organizācijas pārstāv autortiesību un blakustiesību subjektu mantisko tiesību intereses attiecībā uz:

1) publisku izpildījumu, ja tas notiek izklaides vietās, kafejnīcās, veikalos, viesnīcās un citās tamlīdzīgās vietās;
2) nomu, īri un publisku patapināšanu (izņemot datorprogrammas, datu bāzes un mākslas darbus);
3) retranslēšanu pa kabeļiem (izņemot raidorganizāciju tiesības neatkarīgi no tā, vai tās ir pašas raidorganizācijas tiesības vai tās raidorganizācijai nodevuši autortiesību vai blakustiesību subjekti);
4) reproducēšanu personiskām vajadzībām;
5) reprogrāfisku reproducēšanu personiskām vai dienesta vajadzībām;
6) vizuālās mākslas oriģināldarbu tālākpārdošanu;
7) komerciālos nolūkos publicētu fonogrammu izmantošanu.

Mantisko tiesību kolektīvā pārvaldījuma organizācijas saskaņā ar autortiesību un blakustiesību subjektu pilnvarojuma līgumiem pārstāv viņu tiesības un likumīgās intereses visās attiecībās ar jebkuru publisko vai privāto tiesību subjektu, arī tiesās un visos jautājumos, kas attiecas uz šāda veida darbību. Tā kā mantisko tiesību kolektīvā pārvaldījuma organizācijas ir autortiesību un blakustiesību pārņēmēji, tad atkarība no noslēgtā līguma ar autortiesību vai blakustiesību subjektu organizācijas veicamās funkcijas var būt vai nu paplašinātas vai sašaurinātas. Autortiesību likums nosaka tās obligātās funkcijas, ko mantisko tiesību kolektīvā pārvaldījuma organizācijai jāpilda. Ar Latviešu valodas korpusu saistītās tiesiskās attiecības skar šādas organizācijas funkcijas:

1) organizācija vienojas ar darbu un blakustiesību objektu izmantotājiem par atlīdzības lielumu, maksāšanas kārtību un citiem noteikumiem, ar kādiem izdod licences;
2) organizācija izsniedz darbu un blakustiesību objektu izmantotājiem licences to tiesību īstenošanai, ar kuru pārvaldījumu nodarbojas attiecīgās organizācijas, un iekasē licencē paredzēto atlīdzību;
3) organizācija nosaka taisnīgu atlīdzību gadījumos, kad organizācijai ir pienākums administrēt autortiesību un blakustiesību subjektu mantiskās tiesības uz likuma pamata, un iekasē noteikto atlīdzību.

Parasti konkrētu autora darbu administrē konkrēta mantisko tiesību kolektīvā pārvaldījuma organizācija. Nav iespējama situācija, kad darba administrēšanu veic vairākas organizācijas. Lai veicinātu starpvalstu sadarbību dažādu valstu mantisko tiesību kolektīvā pārvaldījumā, organizācijas ar pārstāvniecības līgumu palīdzību nodod savu aizsargājamo objektu pārstāvību konkrētā valstī attiecīgās valsts mantisko tiesību kolektīvā pārvaldījuma organizācijām. Šāda sistēma atvieglo gan autortiesību un blakustiesību subjektu interešu aizstāvības efektivitāti, gan atvieglo potenciālo darbu izmantotāju iespējas saņemt atļauju ārvalstu darbu izmantošanai.

AKKA/LAA ir bezpeļņas organizācija, kas izveidota 2004. gada 2. septembrī Autortiesību un komunicēšanas konsultāciju aģentūras/Latvijas Autortiesību aģentūras reorganizācijas rezultātā, lai kolektīvi pārvaldītu autortiesību īpašnieku mantiskās tiesības. AKKA/LAA pārstāv 2949 Latvijas autorus, kā arī 93 ārvalstu autortiesību organizāciju biedrus. AKKA/LAA ir CISAC (*International Confederation of Societies of Authors and Composers*) biedre. 2005. gada 1. martā izsniegta darbības atļauja veikt mantisko tiesību kolektīvo pārvaldījumu[120]. Latviešu valodas korpusa gadījumā veic mantisko tiesību kolektīvo pārvaldījumu attiecībā uz darba padarīšanu pieejamu sabiedrībai pa vadiem vai citādā veidā, ka tiem var piekļūt individuāli izraudzītā vietā un individuāli izraudzītā laikā. Taču jāņem vēra, ka likumā ietvertās izmaiņas par darbu elektronisko formu izmantošanu ir ieviestas salīdzinoši nesen – ar 2000. gada aprīli, tāpēc autoru iepriekš noslēgtie līgumi ar AKKA/LAA neparedz mantisko tiesību pārstāvību gadījumos, kad darbs tiek izmantots elektroniskā formā internetā, ja vien iepriekš noslēgtais līgums nav papildināts.

LaIPA ir sabiedriska organizācija, kas dibināta 1999. gadā. LaIPA administrē izpildītāju un fonogrammu producentu tiesības Latvijā. LaIPA mērķi ir veikt izpildītāju un producentu tiesību aizsardzības sistēmas sakārtošanu Latvijā, veikt Latvijas izpildītāju un producentu mantisko tiesību kolektīvo pārvaldījumu Latvijā un ārvalstīs, kā arī ārvalstu izpildītāju un producentu mantisko tiesību kolektīvo pārvaldījumu Latvijā. LaIPA pārstāv izpildītāju un producentu intereses saskarsmē ar fiksēto

[120] http://www.km.gov.lv/UI/main.asp?id=17626 – skatīts 22.07.2005.

izpildījumu, fonogrammu un audiovizuālo darbu izmantotājiem. LaIPA pārstāv 260 Latvijas izpildītājus un 27 Latvijas producentus, kā arī 74250 ārvalstu izpildītājus un 6609 ārvalstu producentus. LaIPA ir SCAPR (*The Societies' Council for Collective Management of Performers' Rights*) biedre. 2004. gada 21. jūlijā izsniegta atļauja veikt izpildītāju un fonogrammu producentu mantisko tiesību kolektīvo pārvaldījumu par fonogrammās fiksēto izpildījumu un audiovizuālos darbos fiksēto muzikālo izpildījumu attiecībā uz raidīšanu, ieskaitot vienlaicīgu raidīšanu elektronisko sakaru tīklā un raidīšanu tikai elektronisko sakaru tīklā[121]. Pašlaik LaIPA tikai plāno pārstāvēt tādus blakustiesību darbus, kas nav saistīti ar muzikālu izpildījumu (piemēram, pasaku, lugu, runu un dialogu ieraksti), tāpēc vienīgais risinājums Latviešu valodas runas korpusa darbu apkopošanai ir iespēja vienoties ar blakustiesību darbu autoriem individuāli. LaIPAi izstrādes stadijā šobrīd atrodas vienotu atlīdzības tarifu izveide blakustiesību subjektiem gadījumos, kad darbs tiek izmantots elektronisko sakaru tīklā, tas nozīmē arī interneta tīklā.

6.12. Juridiskie nosacījumi Latviešu valodas korpusa lietotājiem

Tāpat kā ar licences līgumu vai licenci tiek regulētas tiesiskās attiecības, kas rodas starp darba izmantotāju un autortiesību vai blakustiesību subjektu, ir jāregulē tiesiskās attiecības, kas rodas, izmantojot darbus ar Latviešu valodas korpusa starpniecību. Latviešu valodas korpusa uzturētājiem ir jārūpējas par izmantoto darbu aizsardzību no neatļautas to izmantošanas. Šī aizsardzība ir nepieciešama gan nekomerciāla, gan komerciāla korpusa izveides gadījumā. Tiesiskās attiecības ar Latviešu valodas korpusa lietotājiem ir iespējams regulēt šādos veidos:

- ja atsevišķa Latviešu valodas korpusa daļa ir brīvi pieejama kā izmēģinājuma versija, tad pirms tās izmantošanas ir jānodrošina, ka lietotājs piekrīt ievērot korpusa lietošanas noteikumus;
- jebkurā no lietošanas veidiem, slēdzot rakstisku līgumu ar katru lietotāju;
- CD formāta Latviešu valodas korpusa lietošanas noteikumi, tiktu nodrošināti kopā ar CD;
- papildus internetā Latviešu valodas korpusā publicējot korpusa lietošanas noteikumus.

Gan Latviešu valodas korpusa lietošanas noteikumiem, gan lietošanas līgumam ir jāsatur nosacījumi, kas attiecas uz:

- korpusa izmantošanas mērķi;
- autortiesību un blakustiesību ievērošanu darbiem, kas pieejami korpusā;

[121] Ibid.

– autortiesību un *sui generis* tiesību ievērošanu, kas attiecas uz korpusa izveidotājiem.

Tā kā Latviešu valodas korpuss tiks uzturēts Latvijas teritorijā, tad līgums un korpusa lietošanas noteikumi ar korpusa lietotājiem jāveido, ievērojot Latvijas normatīvos aktus. Ja darba izmantošanas atļaujā paredzēts kāds specifisks nosacījums, kas skar arī korpusa lietotājus, tad attiecībā uz konkrēto darbu noteikumos būtu jāatzīmē šis autortiesību vai blakustiesību subjekta noteiktais nosacījums.

Autortiesību likuma 58. pants nosaka datu bāzes izmantotāja tiesības un pienākumus.

„(1) Likumīgam publiski pieejamas datu bāzes izmantotājam ir tiesības iegūt vai atkārtoti izmantot jebkādā nolūkā nebūtisku kvalitatīvi vai kvantitatīvi novērtējamu datu bāzes satura daļu. Šis nosacījums attiecas tikai uz to datu bāzes daļu, kuru likumīgam izmantotājam ir atļauts iegūt vai atkārtoti izmantot.

(2) Likumīgam publiski pieejamas datu bāzes izmantotājam ir jāievēro ar datu bāzē iekļautajiem darbiem vai materiāliem saistīto autortiesību vai blakustiesību subjektu tiesības.

(3) Publiski pieejamas datu bāzes izmantotājs nedrīkst veikt darbības, kas ir pretrunā ar normālu šīs datu bāzes izmantošanu vai kas nepamatoti aizskar šīs datu bāzes veidotāja likumīgās intereses.” Likums nosaka arī gadījumus, kad Latviešu valodas korpusa uzturētājs nedrīkstētu iejaukties likumīga datu bāzes lietotāja darbībās, proti, „likumīgie datu bāzes izmantotāji drīkst:

1) iegūt būtisku datu bāzes satura daļu izglītības vai zinātniskās pētniecības nolūkos, obligāti norādot avotu, turklāt tikai tādā apjomā, kāds nepieciešams nekomerciāla mērķa sasniegšanai;
2) iegūt vai atkārtoti izmantot būtisku datu bāzes satura daļu valsts drošības nolūkos, kā arī administratīviem vai tiesvedības mērķiem.”

Tiesiskā regulācija izvirza arī prasības tehnoloģiskajiem risinājumiem. No tehnoloģiskā viedokļa ieinteresēto personu tiesību aizsardzībai ir jāaplūko šādi aspekti:

1) **valodas korpusa tehnoloģiskā aizsardzība** – autorizācija, autentifikācija, pretkopēšanas pasākumi, tekstu apstrādes iespējas;
2) **lietotāju diferenciācija** – lietotāju dalījums grupās pēc dažādām pazīmēm, piemēram, akadēmiskie lietotāji, komerclietotāji, pilnas pieejas lietotāji, ierobežotas pieejas lietotāji.

6.13. Datu bāzes aizsargāšana

Latviešu valodas korpuss atbilst normatīvajos aktos noteikto pazīmju kopumam, kas nepieciešams, lai Latviešu valodas korpuss tiktu uzskatīts par datu bāzi. Latviešu

valodas korpuss ir neatkarīgu darbu, datu un citu materiālu krājums, kas sakārtots sistemātiski un metodiski, turklāt ir individuāli pieejams elektroniskā vai citādā veidā. Šīs datu bāzes izveidei ir nepieciešams atbilstošs programmnodrošinājums. Atbilstošo programmnodrošinājumu ir iespējams izstrādāt no jauna vai izmantot jau esošus risinājumus. No autortiesību viedokļa tas ir svarīgi tik tālu, cik tas skar attiecīgā programmnodrošinājuma autortiesību ievērošanu. Jāatzīmē, ka tieši datorprogrammas ir tās, kas netiek pakļautas datu bāzu speciālajai autortiesību aizsardzībai, jo ir tikai līdzeklis datu bāzes izveides un uzturēšanas procesā. Datorprogrammas līdzīgi kā literārie darbi ir autortiesību aizsargātas, tāpēc datu bāzes izstrādei ir jāizmanto vai nu licencēta programmatūra, vai arī, ja tiek radīti jauni risinājumi, tie būtu jāaizsargā ar autortiesībām, attiecīgi norādot autortiesību īpašnieku.

1991. gada 14. maija Direktīva 91/250/EEK par datorprogrammu juridisku aizsardzību nosaka speciālu autortiesību režīmu. Lai nodrošinātu datorprogrammu izmantošanas iespēju, direktīva nosaka, ka visas darba ņēmēja ekonomiskās tiesības pāriet darba devējam, ja vien darba līgums nenosaka pretēji. Šī automātiskā autortiesību pāreja ir piemērojama datorprogrammām, ko darba ņēmējs radījis, pildot savus darba pienākumus vai pildot darba devēja norādījumus. Visos pārējos gadījumos ir nepieciešama autortiesību pārejas atruna līgumā. Eiropas Savienības direktīva ierobežo arī darba ņēmēja – autora personiskās tiesības, lai nodrošinātu datorprogrammu izmantošanas iespēju: autors nevar aizliegt darba modificēšanu, tas iespējams tikai gadījumos, kad darba modificēšana aizskar autora cieņu vai reputāciju[122]. Iespējama arī autortiesību pāreja pašam autoram, to var nodrošināt ar attiecīgu līguma atrunu.

Direktīvas nostādnes realizētas Latvijas Republikas Autortiesību likumā. Tā kā paredzams, ka Latviešu valodas korpusa izstrāde notiks Latvijā, tad attiecīgi arī no jauna izstrādātie risinājumi būs aizsargāti tieši ar Latvijas normatīvajiem aktiem. Autortiesību likuma 12. panta 2. daļa nosaka, ja datorprogrammu izstrādājis darbinieks, pildot darba uzdevumu, visas šādā veidā radītās datorprogrammas autora mantiskās tiesības pieder darba devējam, ja vien līgumā nav paredzēts citādi. Jāņem vērā arī direktīvas skaidrojumi par personisko tiesību apjomu, kas nepāriet darba devējam. Ja tiek izmantotas jau izstrādātas datorprogrammas, tām jābūt licencētām, tas nozīmē – datorprogrammām ir jābūt ar legālu izcelsmi.

Pēc tam, kad izstrādāta datu bāzes struktūra un datu bāze piepildīta ar saturu, ir iespējams izdalīt vairāku līmeņu intelektuālā īpašuma aizsardzību:

- no jauna izstrādāto datorprogrammu aizsardzība ar autortiesībām;
- datu bāzes struktūras aizsardzība ar autortiesībām;
- datu bāzes satura aizsardzība ar *sui generis* tiesībām;

[122] http://www.ipr-helpdesk.org/controlador.jsp?cuerpo=principal&seccion=guias&guia=guia3&len=en – skatīts 06.06.2005.

– datu bāzē iekļauto darbu jeb satura aizsardzība ar autortiesībām.

Eiropas Savienības 1996. gada 11. marta Direktīva 96/9/EK par datubāzu tiesisko aizsardzību radīja divpusēju datu bāzes aizsardzību: struktūras aizsardzība ar autortiesībām; un satura aizsardzība ar *sui generis* tiesībām (neatkarīgi no iespējamā satura aizsardzības, ja saturs ir oriģināldarbi). Datu bāze ir definēta kā neatkarīgu darbu, datu vai citu materiālu individuāli elektroniski vai citādi pieejams krājums, kas sakārtots sistemātiski vai metodiski.

Datu bāzes struktūra (datu klasifikācija, datu izvēles iespēja u. tml.) ir autortiesību aizsargāta tik tālu, cik tā ir oriģināla. Šī direktīva neparedz automātisku autortiesību pāreju darba devējam. Tomēr direktīva atļauj dalībvalstīm nodrošināt šādu automātisku autortiesību pāreju, tik tālu, cik tās attiecas uz datu bāzi, kas radīta, pildot darba līguma nosacījumus vai ievērojot darba devēja norādījumus. Lielākā daļa Eiropas Savienības valstu ir nodrošinājušas šādu autortiesību pāreju darba devējam. Dažās valstīs šis risinājums nav atrodams speciālajās datu bāzu normās, bet no vispārējām normām par darba ņēmēja radītajiem darbiem[123].

Svarīgi ir norādīt, ka datu bāzes saturs, ja tas sastāv no oriģināldarbiem, ir aizsargāts arī ar autortiesībām.

Ja datu bāzes izveidošanai nepieciešams būtisks ieguldījums (finansiāls vai darba un patērētā laika ziņā), kvalitatīvi vai kvantitatīvi, tad persona, kas ieguldījusi darbu vai līdzekļus, būs tā, kam pieder *sui generis* tiesības, neskatoties uz to, ka datu bāzes saturs un/vai struktūra nav oriģināli. *Sui generis* tiesības nodrošina datu bāzes radītāju no viņa tiesību aizskāruma, kas var izpausties kā visa datu bāzes satura vai kvalitatīvi vai kvantitatīvi būtiskas daļas iegūšana[124].

Vienīgā norāde, kas dota direktīvā un attiecināma uz *sui generis* tiesībām – „būtisks" ieguldījums – ir fakts, ka „būtisks" ieguldījums var tikt veikts, lai iegūtu, apliecinātu vai piedāvātu datu bāzes saturu[125]. Tas nozīmē, ka dažādās valstīs šis kritērijs var tikt tulkots atšķirīgi, tāpēc nepieciešams būtu iepazīties ar attiecīgas valsts tiesu praksi *sui generis* lietās. Tā kā personiskās tiesības nav atsavināmas, tad mantisko tiesību pāreja no darba ņēmēja darba devējam neatbrīvo no atļaujas saņemšanas no darba autora, ja ir plānots cits darba izmantošanas veids vai ir nepieciešama darba modificēšana. Ja dalībvalstu likumi neparedz automātisku autortiesību pāreju darba devējam, tad autortiesību pāreju var nodrošināt, noslēdzot attiecīgu līgumu.

Latvijas Autortiesību likuma IX nodaļa attiecas uz datu bāzes aizsardzību ar *sui generis* tiesībām. Likums norāda, ka par datu bāzes veidotāju ir atzīstama fiziska vai juridiska persona, kura datu bāzes veidošanā uzņēmusies iniciatīvu un ieguldījuma

[123] Ibid.
[124] Ibid.
[125] Ibid.

risku, turklāt datu bāzes izveidošanā, pārbaudē vai noformēšanā ir jābūt ieliktam būtiskam kvalitatīvam vai kvantitatīvam darbam. Kā norāda likuma 5. panta 2. daļa, šis ieguldījums var būt ne tikai jaunrades rezultāts, bet arī atvasināts darbs, kuru neatkarīgi no oriģināldarba aizsargā autortiesības. *Sui generis* tiesības, kā iepriekš minēts, arī Autortiesību likumā nodrošina tiesisko pamatu datu bāzes satura vai būtiskas kvalitatīvi vai kvantitatīvi novērtējamas tā daļas iegūšanu vai/un atkārtotu izmantošanu. Likums arī aizliedz atkārtoti un sistemātiski iegūt vai atkārtoti izmantot nebūtiskas datu bāzes satura daļas, ja tas notiek ar darbībām, kas ir pretrunā ar normālu datu bāzes izmantošanu vai kas nepamatoti aizskar datu bāzes veidotāja likumīgās intereses. Šāds gadījums iespējams, ja datu bāzi izmanto persona, kurai nav piešķirtas datu bāzes lietotāja tiesības.

Datu bāze ar *sui generis* tiesībām tiek aizsargāta 15 gadus no dienas, kad pabeigta datu bāzes izveide. Gadījumā, ja datu bāze kļūst pieejama sabiedrībai pirms tās pabeigšanas, tad 15 gadu aizsardzības termiņu sāk skaitīt no tā gada 1. janvārī, kas seko pēc dienas, kad datu bāze ir kļuvusi publiski pieejama. Līdzīgi termiņu sāk skaitīt arī pirmajā gadījumā. Tā kā Latviešu valodas korpuss ir ļoti liela apjoma datu bāze, tad visdrīzāk, tā kļūs pieejama sabiedrībai pirms tās pabeigšanas. Priekš šāda veida datu bāzes 15 gadu aizsardzības termiņš ir vērtējams kā ļoti īss, turklāt pastāv iespēja, ka Latviešu valodas korpuss ik pēc noteikta laika tiks atjaunots un papildināts ar jauniem darbiem. Šajā gadījumā piemērots ir Autortiesību likuma piedāvātais risinājums, kas attiecas uz tiesību aizsardzības termiņa pagarinājumu. Autortiesību likuma 60. panta 3. daļa paredz, ja datu bāzes saturā tiek izdarīti jebkādi būtiski kvalitatīvi vai kvantitatīvi novērtējami grozījumi, kā arī tajā radušās izmaiņas uzkrājušos secīgu papildinājumu, izslēgumu vai grozījumu dēļ un to rezultātā var uzskatīt, ka ir veikts būtisks kvalitatīvi vai kvantitatīvi novērtējams jauns ieguldījums, šādai datu bāzei ir tiesības uz savu aizsardzības termiņu, tādējādi datu bāzes aizsardzības termiņu sāk skaitīt no tā gada 1. janvārī, kas seko pēc dienas, kad datu bāze ir grozīta. Tā kā direktīva nosaka šādus principus:

1) pierādīt datu bāzes izveides pabeigšanas datumu ir datu bāzes veidotāja pienākums;
2) pierādīt to, ka ir kritēriji, pēc kuriem var secināt, ka ievērojamas pārmaiņas datu bāzes saturā uzskatāmas par būtisku jaunu ieguldījumu, ir šāda ieguldījuma rezultātā radušās datu bāzes veidotāja pienākums,

tad datu bāzes veidotājam būtu jānodrošina jebkuras datu bāzes izmaiņas datēšana un uzskaite.

Paredzams, ka Latviešu valodas korpusa datu bāzes veidotājs būs juridiska persona, tāpēc ir svarīgi noskaidrot, kādas valsts likumi aizsargās izstrādāto datu bāzi. Juridiskās personas tiesības tiek atzītas saskaņā ar Autortiesību likumu, ja šī juridiskā persona

izveidota saskaņā ar Latvijas vai citas Eiropas Savienības dalībvalsts normatīvajiem aktiem un tās juridiskā adrese, pārvalde vai galvenā darbības vieta atrodas Eiropas Savienībā. Ja juridiskajai personai Latvijas vai citas Eiropas Savienības dalībvalsts teritorijā ir tikai juridiskā adrese, šīs personas darbībām ir jābūt nepārtraukti saistītām ar Latvijas vai attiecīgās Eiropas Savienības dalībvalsts ekonomiku. Ja datu bāze ir veidota ārpus Latvijas un uz to nevar attiecināt šā panta pirmās un otrās daļas noteikumus, tā ir aizsargājama, pamatojoties uz Latvijai saistošiem starptautiskajiem līgumiem.

Tātad, izstrādājot Latviešu valodas korpusa datu bāzi, būtu jāparedz visa saistīta intelektuālā īpašuma aizsardzība un tiesību ievērošana.

Attiecība uz datorprogrammām jāievēro autortiesību aizsardzība, ja datu bāzes izstrādes laikā radīta jauna datorprogramma vai arī jau esošu licencētu datorprogrammu izmantošana.

Datu bāzē iekļauto darbu aizsardzība jānodrošina ne tikai ar tehniskiem, bet arī ar tiesiskiem līdzekļiem pret nesankcionētu un nelikumīgu darbu izmantošanu. Ar katru datu bāzes lietotāju ir jāslēdz līgums par datu bāzes izmantošanu neatkarīgi no tā, vai datu bāze ir domāta izglītības vai pētniecības vai komerciāliem mērķiem. Līgumus var neslēgt par to datu bāzes daļu, kur pieejami darbi, kam beidzies autortiesību aizsardzības termiņš vai darbi vispār netiek aizsargāti ar autortiesībām. Šo datu bāzes daļu varētu izmantot kā izmēģinājuma versiju pirms līguma noslēgšanas. Papildus jāparedz sadaļa, kurā norādīti datu bāzes lietošanas noteikumi. Tā kā datu bāze ir aizsargāta ar *sui generis* tiesībām, tad papildus būtu jānorāda, kas ir datu bāzes veidotājs un kam pieder autortiesības uz datu bāzi.

6.14. Atbildība par autortiesību pārkāpumiem

Par jebkuru autortiesību vai blakustiesību objektu neatļautu vai neatbilstošu izmantošanu, kas aizskar subjektu personiskās vai mantiskās tiesības var iestāties normatīvajos aktos noteiktā atbildība, tā var būt:

1) ar civiltiesisku raksturu;
2) administratīvā atbildība;
3) krimināatbildība.

Autortiesību likums 69. pantā papildus uzskaita autortiesību un blakustiesību subjektu tiesības:

1) prasīt, lai pārkāpējs atzīst viņu tiesības;
2) aizliegt savu darbu izmantošanu;
3) prasīt, lai pārkāpējs atjauno iepriekšējo stāvokli, kāds bija līdz tiesību pārkāpšanai, un pārtrauc prettiesiskās darbības vai neapdraud radošo darbību;

4) prasīt, lai pārkāpējs atlīdzina zaudējumus, arī negūto peļņu, vai arī prasīt, lai pārkāpējs dod kompensāciju pēc tiesas ieskata;

5) prasīt, lai tiek iznīcināti kontrafakti eksemplāri;

6) prasīt, lai starpnieki, kuru sniegtie pakalpojumi tiek izmantoti, lai pārkāptu autortiesību un blakustiesību subjektu tiesības, vai kuri padara iespējamu šāda pārkāpuma veikšanu, veic attiecīgus pasākumus nolūkā pārtraukt izmantotāju iespējas veikt šādus pārkāpumus. Ja starpnieks neveic attiecīgus pasākumus, autortiesību un blakustiesību subjektam ir tiesības vērsties pret starpnieku.

Vēres

Eisenchitz T., Turner P. [1997], "Rights and Responsibilities in the Digital Age: Problems with Stronger Copyright in an Information Society." – *Journal of Information Science*, 23(3). – pp. 209–223.

Paklone I., Lielkalns A., Sosnovska A., Tola K. [1997], *Autortiesības*. – Rīga: Izdevniecība AGB.

Rozenfelds J. [2004], *Intelektuālais īpašums*. – Rīga: Zvaigzne ABC.

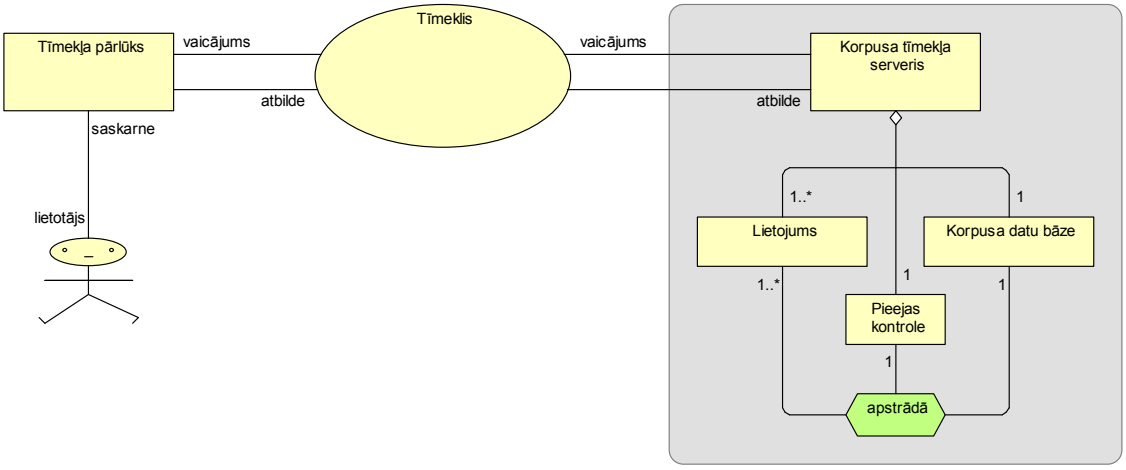
7. Sistēmas uzturēšana un paplašināšana

Korpusa sistēmas vispārējā arhitektūra, programmatūras un datu pilnveides iespējas.

7.1. Korpusa sistēmas vispārējā arhitektūra

Ir iespējami trīs galvenie virzieni:

1. Centralizēta, tīmekļa bāzēta sistēma: visas korpusa daļas un lietojumrīki ir izvietoti vienuviet uz tīmekļa servera, visa saskarne un darbs ar korpusu notiek ar tīmekļa pārlūkprogrammas palīdzību (sk. 7.1. diagrammu).
2. Decentralizēta (galvenokārt attiecībā uz tekstu bāzēm, apakškorpusiem – autortiesības, īpaša pievienotā vērtība, speciālie korpusi, paralēlie korpusi u. tml.), tīmekļa bāzēta sistēma; decentralizācija ir kontrolēta un ierobežota; korpusa komponentu integrācija, savietojamība, programmatūras risinājumu koplietošana; galalietotājiem „no ārpuses izskatās" kā vienota sistēma (sk. 7.2. diagrammu).
3. Decentralizēti, bezsaistes (*off-line*) tekstu krājumi un rīki; komponenšu ieguve: ierobežoti CD izdevumi, lokāli (*desktop*), modulāri korpusa lietojumi; papildus arī tekstu lejupielādes no korpusa servera (iem).

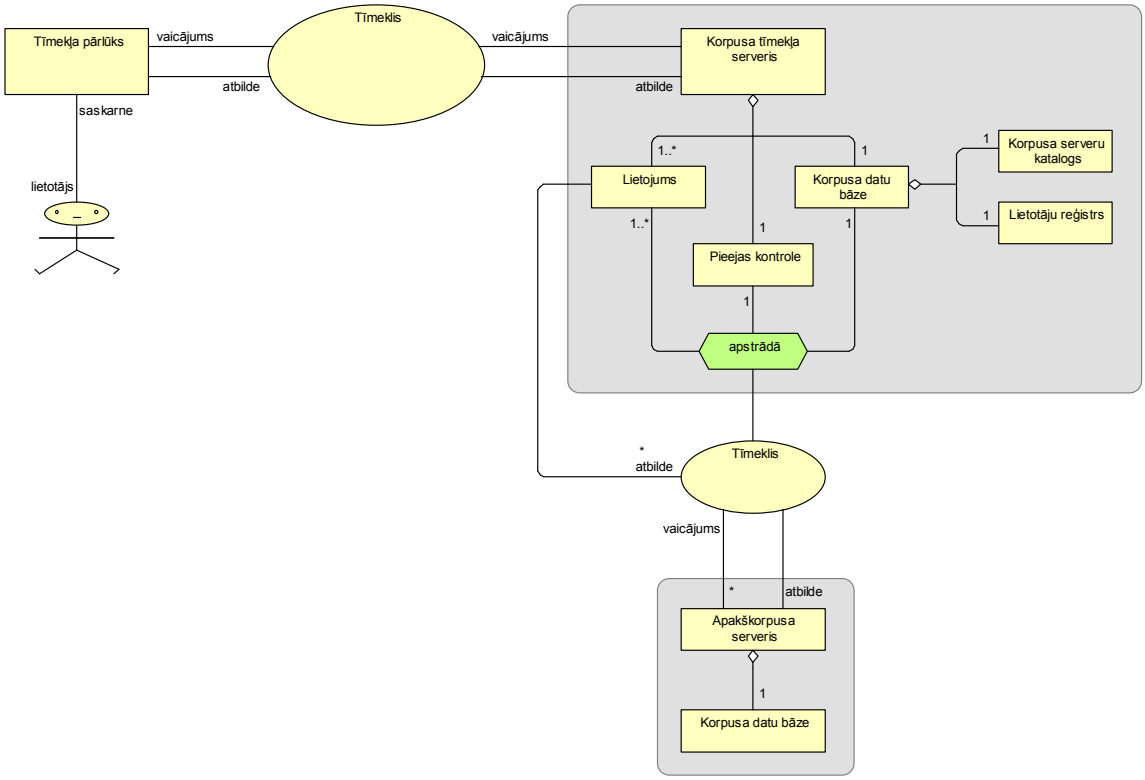


7.1. diagramma – centralizēta, tīmekļa bāzēta arhitektūra (papildus skatīt 5.1. diagrammu).

No izstrādātāju viedokļa piemērotākais risinājums ir tīmekļa bāzēta (klienta-servera) arhitektūra. Būtiskākās valodas korpusa uzturēšanas un attīstības priekšrocības, izvēloties šādu risinājumu, ir:

– iespēja aptvert plašu auditoriju, veicinot korpusa izmantojamību un nozīmīgumu, kā arī veidojot funkcionējošu un vērtīgu atgriezenisko saikni ar korpusa lietotājiem; lietotāju uzvedības, statistikas analīze u. c.;

– ērta sistēmas uzturēšana, izmaiņu ieviešana un attīstīšana, jo viss tekstu korpuss un tam nepieciešamie programmrīki ir centralizēti (vai ierobežoti decentralizēti), tādējādi nav jādomā par sistēmas un marķēto tekstu krājumu aktuālo versiju pieejamību visu kategoriju lietotājiem;
– kontrolēta decentralizācija: ievērojot vienotus savietojamības un korpusa sistēmas moduļu integrēšanas standartus;
– centralizēta pieejas kontroles sistēma un tekstu aizsardzība;
– platformneatkarība gan sistēmas klienta pusē (lietotāju saskarne), gan servera (u) pusē (korpusa lietojumu loģika); servera programmatūrai un izstrādes tehnoloģijām koncepcijas autori iesaka izvēlēties nekomerciālus, platformneatkarīgus, stabilus/veiktspējīgus atvērtā koda risinājumus, piemēram, *Apache*[126] tīmekļa serveris, *MySQL*[127] datubāzes vadības sistēma, *Java*[128] izstrādes vide u. c.



7.2. diagramma – decentralizēta, tīmekļbāzētai arhitektūra (papildus skatīt 5.1. diagrammu).

Dalītas sistēmas gadījumā papildus problemātika, kas jārisina ir komponenšu (lietojumu un marķēto tekstu bāzu) sadarbība: kontrolēta vaicājumu/atbilžu pāradresācija un apstrāde, skaidri komponenšu programmējamie interfeisi, izmantojot,

[126] http://httpd.apache.org – skatīts 25.07.2005.
[127] http://www.mysql.com – skatīts 25.07.2005.
[128] http://java.sun.com – skatīts 25.07.2005.

piemēram, tīmekļa servisu (*web services*) tehnoloģiju, jāuztur sistēmas komponentes formāli aprakstošs reģistrs u. tml.

Savukārt galalietotāju ērtības un priekšrocības tīmekļa darba videi ir šādas:

– plaša pieejamība korpusa saturam un programmrīkiem; darbs var tikt veikts (turpināts) neatkarīgi no atrašanās vietas;
– nepieciešams tikai tīmekļa pieslēgums un pārlūkprogramma, turklāt vajadzīgais programmnodrošinājums ir iegūstams bez maksas (*Microsoft Internet Explorer*, *Mozilla FireFox* u. c.).

Tīmekļa risinājuma priekšrocības un problēmas dažādu robežkritēriju gadījumos uzskatāmāk ir parādītas 7.1. tabulā. Pašreizējā situācijā šādai realizācijai principā ir viens trūkums: ne visiem potenciālajiem lietotājiem vēlamos laikos un vietās ir pieejams tīkla pieslēgums, taču tīmekļa pieejamība un ātrdarbība (no galalietotāju viedokļa) arvien pieaug.

Kritērijs	Priekšrocība	Problēma
Plaša mērķauditorija	**X**	
Atgriezeniskā saikne	**X**	
Korpusa popularitāte (nozīmīgums) un tā plaša izmantojamība	**X**	
Centralizēta sistēma un plaša tekstu bāze	**X**	
Servera (u) resursi un tīkla ātrdarbība		jānovērtē katras operācijas resursu patēriņš; nepieciešama lietotāju pieejas kontroles sistēma
Tehniskās prasības lietotājiem	nepieciešama tikai pārlūkprogramma	
Korpusa uzturēšana un attīstīšana	**X**	
Izstrādes tehnoloģijas	**X**	
Platformneatkarība	**X**	
Īpaši vērtīgi un/vai autortiesību aizsargāti teksti/marķējums	pievienotās vērtības popularizēšana un izmantošanas veicināšana	nepieciešama lietotāju pieejas kontroles sistēma

7.1. tabula – dažādi tīmekļa bāzētas sistēmas aspekti.

Problēmas, kas saistītas ar decentralizētu, lokāla darba sistēmu:

– tekstu un marķējuma papildu aizsardzība (šifrēšana);
– apgrūtināta tekstu un rīku jauninājumu izplatīšana;

- papildprasības klienta sistēmai; iespējama programmrīku platformatkarība, iespējamas saistītas programmatūras licencēšanas problēmas;
- lielā mērā zūd efektīva saikne ar galalietotājiem, lietotāju darba specifikas, īpašību, interešu analīzes iespējas. Šāda analīze būtu noderīga gan valodnieciski pētnieciskiem mērķiem, gan korpusa attīstības virzienu plānošanai.

Pamatā ir jāorientējas uz centralizētu (vai ierobežoti decentralizētu) tīmekļa bāzētu sistēmu. Kā redzams no potenciālo korpusa lietotāju aptaujas rezultātiem, tikai 6% respondentu ir norādījuši, ka nevēlas, lai darbs ar korpusu jebkādā veidā tiktu saistīts ar tīmekli, bet 68% respondentu piekrīt tīmekļbāzētai darba videi.

Atsevišķiem, lietotāju specificētiem tekstu krājumiem (saskaņā ar autortiesībām, arī pievienotās vērtības autoru tiesībām) un klasiskākajiem lietojumrīkiem būtu jābūt pieejamām (vienkāršotām) lejupielādējamām versijām un pēc dažādiem kritērijiem veidotiem CD izdevumiem. Ar standartiem savietojama marķējuma gadījumā brīvi pieejamu tekstu analīzei pastāv iespēja izmantot kādu no pasaulē esošiem, valodneatkarīgiem rīkiem.

7.2. Programmatūras attīstība

- Korpusa un citu latviešu valodas dabīgās apstrādes sistēmu sasaite:
  - tekstu transkribēšana un runas sintēzes funkcionalitāte (lietotāju norādītu tekstu fragmentu izrunāšana) – valodas mācīšanās u. c. nolūkiem; LU MII ir iestrādes šādiem lietojumiem; audio un video sasaiste;
  - leksiska latviešu valodas ontoloģija: zināšanu izguve no tekstu korpusa un mašīnlasāmām vārdnīcām/enciklopēdijām – izmantojamība semantiskajā analīzē/marķēšanā, semantiskas informācijas izguves/konkordanču rīku attīstībā u. c.; konfigurējams semantisko attieksmju izguves rīks leksisko zinību bāzes izveides u. c. vajadzībām (LU MII ir iestrādes);
  - elektronisko vārdnīcu sistēmu un korpusa savstarpējā attīstība un simbioze.
- Lietotāju konti un „darba telpas", dinamiski definējami profili individuālu korpusa skatījumu veidošanai, arī tekstu pievienošana individuālai lietošanai (tīmekļa versijā) – interesējošas esošā korpusa apakškopas atlasīšana, papildus tekstu pievienošana/apstrāde (individuālas „redzamības zonā"), pielāgojama tekstu (kontekstu) un apstrādes rezultātu

vizualizācija, interesējošu marķējuma elementu individuāla izcelšana, meklēšanas u. c. operāciju rezultātu saglabāšana/apvienošana u. tml.

– Tīmekļa „zirneklis" latviešu valodas tīmekļa resursu automatizētai „apstaigāšanai", tekstu vākšanai, turpmākai apstrādei un tīmekļa korpusa veidošanai. Šeit par pamatu var izmantot kādu no Latvijā/pasaulē jau esošajiem, konfigurējamiem „zirnekļu" dzinējiem.
– Pie paplašināšanas var skaitīt arī citu „neklasisko" funkcionalitāti, kas tika minēta 5.3.3. sadaļā, piemēram, lietotāju atgriezeniskās saites uzturēšanu un izmantošanu. Izstrādāto lietojumu uzlabošana/optimizēšana, detalizētu, pētniecības mērķiem specifisku papildfunkciju izstrāde.
– Marķējuma līmeņu un metadatu modeļu paplašināšana.
– Atsevišķu lietojumu kā tīmekļa servisu (*web services*) nodrošināšana korpusa programmatiskai (automātiskai) izmantošanai no citām, neatkarīgām sistēmām, saistītām ar latviešu valodas analīzi/apstrādi; arī decentralizētas sistēmas uzturēšanas un paplašināšanas iespēju nolūkā. Semantiskā tīmekļa tehnoloģiju atbalsts metadatu un lietojumu – servisu – līmenī; īpaši tas ir attiecināms uz latviešu valodas semantikas zinību resursiem.

8. Korpusa izstrādei nepieciešamā izpildes laika plānojums ar pamatojumu

Balstoties uz korpusa koncepciju, autori piedāvā šādu latviešu valodas vispārīgā korpusa izstrādes darba plānu 5–10 gadiem, sadalot to divos etapos – minimālā un maksimālā programma. Pirms sākt korpusa sistēmas izstrādi, ir jāizstrādā detalizēts šīs sistēmas projektējums, ievērojot korpusa koncepcijas ieteikumus. Sistēmas projektējums ir atsevišķs sistēmas izstrādes posms, kas šobrīd netiek piedāvāts.

Jāpiebilst, ka plānojuma izpilde ir atkarīga arī no piešķirto līdzekļu apjoma un cilvēkresursiem.

8.1. Valodas korpusa izveides minimālā programma (5 gadi)

1. gads. Esošo valodas apstrādes rīku un standartu izvērtēšana un sistēmas projektēšana, un datu uzkrāšana.

Latviešu valodas korpusam nepieciešamo programmnodrošinājumu un attiecīgo marķēšanas līmeņu standartu izvērtēšana un izvēle, vajadzības gadījumā – kombinēšana un paplašināšana.

Sistēmas projektēšana: detalizēta arhitektūras modelēšana; datu modeļu definēšana (metadati, marķēšanas gramatikas, formāti, datu bāzu shēmas), ievērojot starptautiskas vadlīnijas; programmatūras komponenšu detalizēta projektēšana un to savietojamības un sadarbības modeļa specificēšana; satura administrēšana un pieejas kontrole; automatizētas tekstu analīzes un marķēšanas metodika (pirmie līmeņi).

Datu uzkrāšana: tiek savākts 1 miljonu vārdlietojumu liels sabalansēts rakstītās valodas korpuss, kura sastāvs tika aprakstīts koncepcijas 1. nodaļā.

Datu uzkrāšana: elektronisko datu apstrāde un jaunu datu skenēšana.

Teksti tiek sagatavoti mašīnlasāmā formā, ņemot vērā projektējumu:

(1) daļēji automatizēta strukturālā marķēšana;

(2) metadatu pievienošana.

Datu pārbaude: ievadītos tekstus nepieciešams pārlasīt un pārbaudīt, marķēt oriģināla iespiedkļūdas un pārbaudīt strukturālo marķējumu (atbilstoši teksta veidam un specificētās gramatikas prasībām).

2. gads. Datu uzkrāšanas turpināšana un strukturāli marķēto datu morfoloģiskā anotēšana. Kopējās korpusa sistēmas karkasa izstrāde. Pirmo lietojumrīku izveide.

Datu uzkrāšanas turpināšana – 1 miljons vārdlietojumu ar strukturālo marķējumu.

Datu uzkrāšana: elektronisko datu apstrāde un jaunu datu skenēšana.

Datu morfoloģiskā marķēšana:

(1) 10 000 vārdlietojumu mēnesī manuāla morfoloģiskā marķēšana, tādējādi gada laikā sasniedzot aptuveni 100 000 vārdlietojumu;

(2) uz manuālās marķēšanas rezultātu pamata tālāk tiktu attīstīti un uzlaboti jau esošie automātiskie morfoloģiskās marķēšanas rīki/iestrādes. Aptaujājot potenciālos korpusa lietotājus, sabiedrība „Tilde" piedāvāja korpusa izstrādē izmantot viņu morfoloģisko analizētāju. Morfoloģiskās analīzes pieredze ir arī LU MII. Izvēloties kādu no esošiem rīkiem, būtu jāveic tā aprobācija un paplašināšana.

Sagaidāms, ka būs jāpapildina morfoloģiskās marķēšanas metodika. Marķēšanas kvalitātes uzlabošanai ir jāveic teorētiskie pētījumi, jāuzlabo principi būtisku problēmjautājumu risināšanai – galvenokārt daudznozīmības novēršanai.

Kopējās korpusa sistēmas karkasa izstrāde – korpusa arhitektūras pamatu implementācija (datubāze un tās saskarne, satura vadība, indeksēšanas mehānisms, tekstu/marķējuma versiju kontrole, pieejas kontroles sistēma, komponenšu (lietojumrīku u. c.) savietojamības un kooperācijas vadība). Pirmo lietojumrīku izveide: navigācija, kontekstu ierobežošana un pozicionēšana, transformēšana prezentācijas formātos; vienkārša meklēšana un statistika. Lietotāju saskarnes pirmais tuvinājums.

3. gads. Datu uzkrāšanas turpināšana un strukturāli marķēto datu automatizēta morfoloģiskā anotēšana. Sintaktiskā marķējuma pievienošana. Lietojumrīku funkcionalitātes uzlabošana.

Datu uzkrāšanas turpināšana – 1 miljons vārdlietojumu ar strukturālo marķējumu. Datu uzkrāšana: elektronisko datu apstrāde un jaunu datu skenēšana.

Esošo strukturāli marķēto datu automatizēta morfoloģiskā marķēšana – 2 miljoni vārdlietojumu.

Automatizēta morfoloģiskā marķēšanas rezultativitāte (progress) ir atkarīga no līdzšinējo marķēšanas principu precizitātes. Nepārtraukti notiek principu precizēšana un uzlabošana.

Sintaktiskā marķēšana – izmantojot morfoloģiski marķēto korpusu, tiek uzsākta teikumu sintaktiskās analīzes principu, metodikas un programmatūras izstrāde. Šajā darbā būtu jāizmanto jau līdzšinējās iestrādes latviešu valodai vai jāmeklē iespējas piemērot citu fleksīvo valodu, piemēram, čehu, krievu valodas sintaktiskos analizatorus. Līdzīgi kā automatizēta morfoloģiskās marķēšanas sistēmas izstrādē, arī tagad ir manuāli jāsamarķē dati. Domājams, ka sākumā vajadzīgas 4000 vienības (teikumi). Tās neatkarīgi ir jāmarķē diviem cilvēkiem, jāveic šo vienību pārbaude un salīdzināšana, dažādo interpretāciju gadījumā jāpanāk vienošanās.

Visām teorētiskajām nostādnēm, kas izmantotas korpusa morfoloģiskajā un sintaktiskajā marķēšanā, jābūt paskaidrotām un publiski pieejamām, lai lietotājs ar tām varētu iepazīties.

Kopējās korpusa sistēmas paplašināšana. Lietojumrīku funkcionalitātes uzlabošana atbilstoši nākamajam marķēšanas līmenim: meklēšana, izmantojot morfoloģisko analīzi, statistika.

4. un 5. gads. Korpusa papildināšana un morfoloģiskās un sintaktiskās anotēšanas pilnveidošana un turpināšana. Korpusa sistēmas pilnveidošana. Lietojumrīku funkcionalitātes uzlabošana.

Datu uzkrāšana – 1. etapa beigās (5 gadu laikā) latviešu valodas korpusā vajadzētu būt 5 – 7 miljoniem vārdu ar strukturālo marķējumu; 5 miljoniem vārdlietojumu ar morfoloģisko marķējumu un 20 000 teikumu ar sintaktisko marķējumu. Datu uzkrāšana: elektronisko datu apstrāde un jaunu datu skenēšana.

Korpusa sistēmas paplašināšana un uzlabošana (izstrādes laikā atklāto trūkumu novēršana). Lietojumrīku funkcionalitātes uzlabošana atbilstoši nākamajam marķēšanas līmenim: sintaktiskie koki, meklēšana, izmantojot sintaktisko analīzi, sintaktiskā statistika.

8.2. Valodas korpusa izveides maksimālā programma (+ 5 gadi)

Izmantojot iepriekšējās fāzes rezultātus, koncepcijas autori piedāvā attīstīt latviešu valodas korpusu. Visticamāk, ka uz šo laiku, attīstoties tekstu nozīmes analīzes un semantiskā tīmekļa tehnoloģijām, būs aktuāli papildu marķēšanas līmeņi. Tāpēc maksimālajā programmā koncepcijas autori uzsvaru liek uz semantisko anotēšanu, kas nav iespējama bez korpusa attīstības minimālās programmas izpildes un ir cieši saistīta ar latviešu valodas teorētiskajiem pētījumiem un konkrētām lingvistisko zināšanu bāzēm. Arī lietotāju aptaujas rezultāti rāda, ka tekstu semantiskā analīze ir ļoti būtiska un nepieciešama.

Šī koncepcija ir veidota tehnoloģiski neitrāla. Šobrīd nav iespējams prognozēt tehnoloģisko risinājumu gaitu nākotnē, jo vienlaicīgi norit pētniecība, kuras rezultātus varēs izmantot arī valodas korpusa semantikas analīzē.

Latviešu valodas korpusa maksimālajā programmā ir jāparedz arī agrāk uzkrāto latviešu valodas tekstu (piem., LU MII rīcībā esošo latviešu literatūras klasikas un folkloras krājumu) pievienošana. Tādējādi mūsdienu latviešu valodas korpusam, kuru veido pirmos piecus gadus, tiktu pievienoti ievērojami tekstu uzkrājumi, kas aizsāktu latviešu valodas diahroniskā korpusa izveidi. Pirms esošo datu pievienošanas, jāveic tekstu pārbaude, metadatu pievienošana un marķēšana atbilstoši izvēlētajam korpusa datu modelim.

8.3. Runas korpusa izstrāde

Koncepcijas autori paredz, ka vispārīgajā latviešu valodas korpusā runātās valodas dati veido 10%. Vajadzētu izveidot 100 000 vārdlietojumu korpusu, no kuriem lielākā daļa būtu dažādi runāti (Saeimas debates, lugas u. c.) teksti, kopā tas būtu 18 stundu garš skaņu ieraksts. Katru gadu runas korpuss tiks papildināts ar 18 stundu garu transkribētu un (strukturāli un morfoloģiski) marķētu runātu tekstu.

Runas korpusa izveides posmi:

1) tekstu uzkrāšana,
2) transkripcija,
3) strukturālā marķēšana,
4) morfoloģiskā marķēšana,
5) lietojumrīku izstrāde.

Tekstu uzkrāšana. Lai varētu runāt par sabalansētu un pilnīgu runas korpusu, tam vajadzētu ietvert gan spontānu, gan iepriekš sagatavotu runu. Nepieciešams runātās latviešu valodas korpusā iekļaut gan monologus, gan dialogus, aptverot gan spontānu, gan sagatavotu runu. Datus iegūst, ieskaņojot runu un digitalizējot jau esošo audiomateriālu.

Transkripcija. Runas transkribēšana ir diskursa atšifrēšana un fiksēšana mašīnlasāmā formā (parasti ortogrāfijā, retāk – fonētiskajā transkripcijā), norādot, piemēram, pauzes, ieelpas un izelpas vietas, runātāju maiņu sarunas laikā.

Runas korpusa strukturālā marķēšana. Vienlaicīgi ar tekstu transkribēšanu ir jāveic runas korpusa strukturālā marķēšana. Vēlams runātās valodas strukturālā marķēšanā izmantot tās pašas pazīmes, kuras tiek lietotas, raksturojot rakstītās valodas tekstus, protams, papildinot ar tieši runas korpusam raksturīgiem elementiem, piem., ziņas par runātājiem (dzimšanas dati, dzimums, runātāju radniecības pakāpe, izglītība u. tml.), kontekstu (sarunas norises vieta, laiks, situācijas apraksts u. tml.), informācija par runas datnes ieskaņošanas laiku, ilgumu.

Morfoloģiskā marķēšana. Transkribētu runu morfoloģiski marķē, izvēloties teksta korpusa morfoloģiskās anotēšanas paņēmienus.

Lietojumrīku izstrāde. Šeit izmantojami (paplašināmi) tekstu korpusa lietojumi, bet papildu ir jāpievieno tieši runas datiem raksturīgie (piem., iespēja noklausīties izvēlētus tekstu fragmentus), meklēt pēc runu raksturojošiem parametriem.

8.4. Paralēlā korpusa izveide

Koncepcijas autori piedāvā izveidot angļu-latviešu paralēlo korpusu, kas sastatīts teikuma līmenī. Paralēlā korpusa izveide atšķiras no vienvalodu korpusa ar to, ka ir nepieciešami papildus programmrīki tieši datu priekšapstrādei, lai tos varētu izmantot tekstu sastatīšanai teikuma līmenī.

Papildus ir jāizstrādā (vai jāpiemēro latviešu valodai) tekstu sastatītājs (teikumu līmenī).

Attiecībā uz lietojumrīkiem ir jāizveido (vai jāizvēlas un jāpiemēro esošās) paralēlās konkordances lietojums.

* Šajā nodaļā netiek apskatīts speciālā korpusa (izlokšņu, studentu (valodas apguvēju)) un paralēlā korpusa veidošanas laika plānojums.

8.5. Laika plānojums

[...]

9. Latviešu valodas korpusa izveidei nepieciešamo izmaksu aprēķins, ņemot vērā piedāvātos risinājumus. Citu Eiropas Savienības valstu pieredze finansiālo jautājumu risināšanā. Iespējas izmantot Eiropas Savienības fondu finansējumu

9.1. Latviešu valodas korpusa izveidei nepieciešamo izmaksu aprēķins, ņemot vērā piedāvātos risinājumus

[...]

9.2. Citu Eiropas Savienības valstu pieredze finansiālo jautājumu risināšanā

Veidojot valodas korpusu, svarīgs ir arī finansiālais aspekts, kam pievērsa uzmanību, meklējot datus par citu valstu pieredzi valodas korpusu izveidē (sk. plašāk 2. nodaļu). Apkopojot informāciju, tika secināts, ka atsevišķi korpusi veidoti plašāku projektu ietvaros un saņēmuši arī Eiropas Savienības finansējumu, piemēram, Īru valodas nacionālais korpuss[129], kas veidots kā Eiropas Savienības projekta PAROLE daļa no 1996. – 1999. gadam, tomēr līdztekus Eiropas Savienības finansējumam šim korpusam piešķirts arī valsts finansējums. Arī dažādi nīderlandiešu valodas resursi[130] ir uzkrāti, pateicoties ES fondu atbalstam (NERC, PP–PAROLE, LE–PAROLE un TELRI).

Liela daļa korpusu tāpat kā iepriekšminētie ir veidoti ar valsts iestāžu finansiālu atbalstu. Piemēram, Britu nacionālo korpusu[131] atbalstījusi Lielbritānijas Zinātnes un tehnikas padome, Britu bibliotēka, Britu akadēmija u. c. Valsts finansiālu atbalstu saņēmis arī Čehu nacionālais korpuss[132], Lietuviešu valodas tekstu korpuss[133], Oslo daudzvalodu korpuss[134], Norvēģu–angļu paralēlais korpuss[135] un Sintaktiski anotētais vācu laikrakstu korpuss[136], Krievu valodas Nacionālajam korpusam[137] ir piešķirts Krievijas Humanitārā zinātniskā fonda grants no 2003. – 2005. gadam, holandiešu valodas korpusu[138] ir atbalstījusi Ekonomikas ministrija, Izglītības, kultūras un zinātnes ministrija un Nīderlandes Zinātniskās izpētes organizācija. Franču valodas gramatikas mācīšanās korpusa izveidei[139] savu atbalstu ir sniegusi universitāte, Franču valodas mācību runātās valodas korpusu[140] – *Research Endowment Trust Fund of the University*

[129] http://www.ite.ie/corpus – skatīts 27.07.2005.
[130] http://www.inl.nl/eng/pub/grancon.htm – skatīts 27.07.2005.
[131] http://www.natcorp.ox.ac.uk – skatīts 27.07.2005.
[132] http://ucnk.ff.cuni.cz/english – skatīts 27.07.2007.
[133] http://donelaitis.vdu.lt – skatīts 27.07.2005.
[134] http://www.hf.uio.no/iba/OMC/English/index_e.html – skatīts 27.07.2005.
[135] http://www.hf.uio.no/iba/prosjekt - skatīts 27.07.2005.
[136] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus – skatīts 27.07.2005.
[137] http://www.ruscorpora.ru – skatīts 27.07.2005.
[138] http://www.onderzoekinformatie.nl/en/oi/nod/onderzoek/OND1306158 – skatīts 27.07.2005.
[139] http://at.its.uiowa.edu/atac/awards/2001/french-corpora.shtml – skatīts 27.07.2005.
[140] http://www.flloc.soton.ac.uk/reading.html – skatīts 27.07.2005.

of Reading, Poļu valodas nacionālo korpusu[141] ir atbalstījusi Polijas Zinātniskās izpētes komiteja. Ungāru valodas resursu[142] uzkrāšanai un apstrādei atbalstu sniedz Izglītības ministrija.

Apkopojot informāciju gan par valodu korpusiem, gan arī par dažādiem uzkrātiem resursiem starptautisko projektu ietvaros, var secināt, ka dažādi nozīmīgi resursi ir tieši uzkrāti šādu projektu ietvaros, piemēram, MULTEXT[143], MULTEXT–EAST[144], ELRA[145], Elsnet, SOCRATES, kā arī citu ES projektu ietvaros[146].

Jāsecina, ka daži valodas korpusi ir ieguvuši arī finansiālu struktūru atbalstu, piemēram, Čehu nacionālo korpusu ir atbalstījuši Čehijas Nacionālā banka un Komercbanka.

Runājot plašāk – ne tikai no ES valstu viedokļa, var minēt, ka ASV runātās valodas korpusus[147] galvenokārt atbalsta Aizsardzības ministrija. Tāpat arī izpēti atbalsta dažādi fondi un iestādes, bet, piemēram, Ķīnas runas valodas korpusu atbalsta Ķīnas Zinātnes fonds.

Tādējādi var secināt, ka valodas korpusu izveidi pārsvarā finansē dažādas valsts iestādes dažādu projektu ietvaros, kā arī reizēm finanšu iestādes, piemēram, Čehu nacionālā korpusa gadījumā, tāpat arī dažu korpusu izveidi finansē ES fondi. Nav mazsvarīgs fakts, ka dažādu ES projektu ietvaros dažādām valodām tika uzkrāti nozīmīgi resursi, tai skaitā arī latviešu valodai (par to sk. plašāk 1. nodaļā), kas viennozīmīgi ir atbalsts valodas korpusa izveidē, tomēr tas nebūt neveido valodas korpusu. Problēmu rada tas, ka daudzām valstīm valodas tehnoloģijas ir attīstītas tālāk, tāpēc mums ir sarežģīti pieteikties daudziem projektiem, jo nav izveidots izstrādāts pirmais etaps.

9.3. Iespējas izmantot Eiropas Savienības fondu finansējumu

No ES fondiem varētu tikt izmantots Eiropas Reģionālās attīstības fonds[148]. Domājams, ka ar tā palīdzību varētu uzkrāt resursus un domāt par valodas korpusa izveidi, izmantojot „Informācijas un sakaru tehnoloģijas attīstības" aktivitāti vai arī „Atbalstu lietišķās zinātnes attīstībai valsts zinātniskajās institūcijās" aktivitāti. Tāpat

[141] http://www.utexas.edu/world/sls/longann.htm – skatīts 27.07.2005.
[142] http://216.239.59.104/search?q=cache:w7135b1Swg8J:www.coli.uni-saarland.de/conf/linc-04/csendes.pdf+Language+corpus+is+funded+by&hl=lv – skatīts 27.07.2005.
[143] http://www.lpl.univ-aix.fr/projects/multext/index.html – skatīts 27.07.2005.
[144] http://www.lpl.univ-aix.fr/projects/multext-east – skatīts 27.07.2005.
[145] http://www.elra.info - skatīts 27.07.2005.
[146] http://www.inl.nl/eng/europe/projects.htm – skatīts 27.07.2005.
[147] http://cslu.cse.ogi.edu/HLTsurvey/ch12node5.html – skatīts 27.07.2005.
[148] ERAF fonds http://www.cfla.gov.lv/?sadala=3 – skatīts 27.07.2005.

arī būtu izmantojams Eiropas Sociālais fonds[149] resursu ieguvei, uzkrāšanai u. c. ar valodas korpusa izveidi saistītām darbībām, piemēram, šādās aktivitātēs:

3.2.1. aktivitāte „Izglītības programmu uzlabošana sākotnējā profesionālajā izglītībā ekonomikai svarīgās nozarēs";

3.2.3.2. aktivitāte „Studiju programmu īstenošana un studiju procesa kvalitātes uzlabošana dabaszinātņu un tehnoloģiju ietilpīgās nozarēs";

3.2.4.2. aktivitāte „Tālākizglītības iespēju paplašināšana ekonomikai svarīgās nozarēs";

3.2.5. aktivitāte „Atbalsts akadēmiskā personāla un pedagogu tālākizglītībai";

3.2.6.2. aktivitāte „Atbalsts akadēmiskā personāla un pedagogu tālākizglītībai";

3.2.6.3. aktivitāte „Atbalsts mācību prakses īstenošanai profesionālās izglītības un augstākās izglītības studentiem";

3.2.7.2. aktivitāte „Profesionālās orientācijas un konsultēšanas pasākumi izglītības iestādēs".

Ir arī pieejamas dažādas citas programmas, piemēram, *Culture 2000*[150], kas atbalsta dažādus kultūras projektus, *Lingua*[151], kuras ietvaros var saņemt līdzekļus mācību līdzekļu un materiālu izveidei, ņemot vērā programmas mērķus, kā arī citas programmas.

Tomēr jāsecina, ka nevienā no tām nevar tieši pieteikt projektu valodas korpusa izveidei, bet gan var pieteikt projektus resursu izveidei, ieguvei, metodikai, mācību kursu izveidei; galaprodukti varētu būt noderīgi valodas korpusa izveidei. Principā viss ir atkarīgs no konkrēta izsludinātā projekta. Jāsecina, ka sākumā būtu izveidojama valodas korpusa bāze, lai varētu no Eiropas fondiem piesaistīt līdzekļus tālākai tās attīstībai.

No 2000. gada ir aizsākta programma *E–content*[152], kas atbalsta daudzvalodu satura izveidi inovatīviem, tiešsaistes pakalpojumiem visā Eiropas Savienībā, digitāla satura pieejamību, lietošanu un izmantošanu. Programma arī atbalsta ģeogrāfiska satura, izglītības, kultūras, zinātnes un akadēmiska satura izveidi[153].

Pagaidām nav izsludināts projektu konkurss, bet ir iespējams izlasīt programmas saturu, tai skaitā latviešu valodā[154], lai uzzinātu programmas mērķus un iespējas piedalīties projektu konkursā. Iepriekšminētā Lēmuma latviskajā versijā ir arī atrodams paredzamais izdevumu sadalījums, piem., 40-50% digitālā satura pieejamības,

[149] ESP http://www.esflatvija.lv/index.php?selected=10&lang=1 – skatīts 27.07.2005.
[150] http://www.km.gov.lv/UI/main.asp?id=14797 – skatīts 27.07.2005.
[151] http://www.km.gov.lv/UI/imagebinary.asp?imageid=1643 – skatīts 27.07.2005.
[152] http://www.cordis.lu/econtent/home.html – skatīts 27.07.2005.
[153] http://europa.eu.int/information_society/activities/econtentplus/index_en.htm – skatīts 27.07.2005.
[154] http://europa.eu.int/information_society/activities/econtentplus/docs/prog_decision_2005/econtentplus_decision_lv.pdf – skatīts 27.07.2005.

lietošanas un izmantošanas veicināšanai Kopienas līmenī, 45–55% kvalitātes uzlabošanas veicināšana un labākas prakses veicināšana saistībā ar digitālo saturu starp satura nodrošinātājiem un lietotājiem, un starp nozarēm, un 8–12% digitālā saturā ieinteresēto pušu sadarbības un informētības veicināšanu. Iespējams, ka šīs programmas resursus varētu piesaistīt, lai attīstītu valodas korpusu.