(Ministru kabineta  
201\_.gada \_\_\_.\_\_\_\_\_\_\_\_  
rīkojums Nr.\_\_\_)

ERAF projekta   
„Daudzvalodu korpusa un mašīntulkošanas infrastruktūras izveide e-pakalpojumu pieejamības nodrošināšanai”   
Sistēmas darbības koncepcijas apraksts

Rīga 2011

Īss satura izklāsts

Kritiski svarīgs e-Pārvaldes uzdevums ir nodrošināt e-pakalpojumu pieejamību visām iedzīvotāju grupām neatkarīgi no to valodu zināšanām, jo tie ir paredzēti izmantošanai gan Latvijas pilsoņiem, gan nepilsoņiem, gan ārzemniekiem. Pašreiz vairums e-pakalpojumu ir izmantojami tikai latviešu valodā. Visu e-pakalpojumu nodrošināšana vairākās valodās ievērojami sadārdzinātu un palēninātu to izstrādi un uzturēšanu. Pastāvošo valodas barjeru e-pakalpojumu pieejamībā var efektīvi risināt, ieviešot mašīntulkošanas (MT) iespēju e‑pakalpojumos: izstrādājot attiecīgus risinājumus un ieviešot tos jau esošajos un plānotajos e-pakalpojumos, kā arī uz vienotas MT sistēmas bāzes izstrādājot atsevišķus lietojumus vai risinājumus valsts pārvaldes elektroniskajos resursos esošās informācijas tulkošanai latviešu-angļu, angļu-latviešu un latviešu-krievu valodās.

Būtiski ir nodrošināt tieši aktuālās informācijas pieejamību lietotājiem saprotamā valodā, gadījumos, kad šādas informācijas tulkošanu izmantojot tulkus nav iespējams operatīvi nodrošināt (piemēram, paziņojumi par pārtraukumiem elektroapgādē vai rīcībai noteiktos gadījumos utt.).

Latvijas iedzīvotājiem un viesiem, kas nepārvalda latviešu valodu, latviešu-angļu un latviešu-krievu MT iespēja ļaus brīvi piekļūt informācijai, kas savukārt veicinās ārvalstu studentu, uzņēmēju, izglītības, zinātnes un kultūras darbinieku integrāciju Latvijas sabiedrībā.

Papildus tam angļu-latviešu MT serviss Latvijas iedzīvotājiem atvieglos pieeju citu ES valstu informācijai, palīdzēs nodrošināt to tiesību aizsardzību, strādājot vai mācoties citās ES dalībvalstīs. Latviešu-krievu MT iespēja iestāžu elektroniskajos informācijas resursos mazinās informatīvo telpu nošķirtību, kura nodala to krieviski runājošo iedzīvotāju daļu, kas nepietiekoši pārvalda latviešu valodu. Līdz ar to MT infrastruktūra latviešu-angļu, angļu-latviešu un latviešu-krievu valodām ir prioritāra, bet vēlāk to var papildināt arī ar citām valodām.

Informācijas un satura pakalpojumu apjomam nepārtraukti pieaugot, tulkotāju piesaiste plaša apjoma tulkošanai kļūst nesamērīgi dārga, tāpēc pasaulē un ES arvien plašāk sāk izmantot MT tehnoloģijas. Eiropas Komisija nodrošina MT servisu valsts pārvaldes iestādēm, diemžēl tas ir pieejams tikai lielajām Eiropas valodām[[1]](#footnote-2). Vairākās ES valstīs MT servisi publisko pakalpojumu nodrošināšanai tiek izstrādāti un nodrošināti arī mazākām valodām, piemēram, basku-spāņu[[2]](#footnote-3) vai angļu-lietuviešu[[3]](#footnote-4) mašīntulkošanai.

Šādas sistēmas izveides projektam būs nozīmīga loma latviešu valodas tehnoloģiskās attīstības nodrošināšanā, valodas resursu bāzes izveidē tālākām akadēmiskām un lietišķām izstrādēm un integrācijai kopējā ES valodas resursu infrastruktūrā.

Projekta rezultātā Latvijas e-Pārvaldes infrastruktūra tiks papildināta ar MT iespēju, ko varēs integrēt jau esošajos un plānotajos e‑pakalpojumos. MT tiek plānots integrēt citu e‑pakalpojumu interfeisa vai loģikas līmeņos, tādējādi ļaujot pakalpojuma lietotājam ātri un ērti tulkot visu pakalpojuma sniegto informāciju angļu, krievu un, iespējams, arī citās valodās. Lai MT iespēja būtu viegli integrējama, tai ir jāizveido piemērots interfeiss (*API*), lai tā būtu viegli izmantojama visos e-pakalpojumos un valsts pārvaldes informācijas sistēmās.

Saturs

[1 Par dokumentu 6](#_Toc297232460)

[1.1 Izmantotie termini, akronīmi un saīsinājumi 6](#_Toc297232461)

[2 Esošās situācijas analīze 8](#_Toc297232462)

[2.1 Esošie procesi 8](#_Toc297232463)

[2.2 Esošajos procesos iesaistītās personas 9](#_Toc297232464)

[2.3 Esošā sistēma 10](#_Toc297232465)

[2.3.1 Esošie E-pakalpojumi un to tehniskie risinājumi 10](#_Toc297232466)

[2.4 Esošais tiesiskais regulējums 11](#_Toc297232467)

[2.5 Esošie politikas dokumenti 12](#_Toc297232468)

[2.6 Mūsdienīgu mašīntulkošanas risinājumu analīze 13](#_Toc297232469)

[2.6.1 MT attīstības vēsture 13](#_Toc297232470)

[2.6.2 Esošie MT risinājumi 13](#_Toc297232471)

[2.6.3 Mašīntulkošanā izmantotās metodes 14](#_Toc297232472)

[3 Izmaiņu pamatojums un būtība 16](#_Toc297232473)

[4 Plānotās situācijas apraksts 17](#_Toc297232474)

[4.1 Plānotie procesi 17](#_Toc297232475)

[4.1.1 Analīze 17](#_Toc297232476)

[4.1.2 Esošo valodas korpusu izpēte 19](#_Toc297232477)

[4.1.3 Valodas Korpusu izveide 24](#_Toc297232478)

[4.1.4 MT sistēma 26](#_Toc297232479)

[4.1.5 Infrastruktūra 28](#_Toc297232480)

[4.1.6 Integrēšana citos e-pakalpojumos 28](#_Toc297232481)

[4.1.7 Kvalitātes mērījumi 29](#_Toc297232482)

[4.2 Plānotajos procesos iesaistītās personas 29](#_Toc297232483)

[4.3 Plānotā sistēma 31](#_Toc297232484)

[4.3.1 MT sistēma 31](#_Toc297232485)

[4.3.2 Infrastruktūra 33](#_Toc297232486)

[4.3.3 Atbalsta stratēģija 36](#_Toc297232487)

[4.4 Plānotais tiesiskais regulējums 36](#_Toc297232488)

[4.5 Plānotie politikas dokumenti 38](#_Toc297232489)

[5 Plānotās sistēmas konceptuālā risinājuma analīzes kopsavilkums 38](#_Toc297232490)

[5.1 Citu valstu pieredzes analīze 38](#_Toc297232491)

[5.2 Analīze par mašīntulkošanas servisu integrāciju e-pārvaldes infrastruktūrā 38](#_Toc297232492)

[5.3 Kopsavilkums par plānoto informācijas sistēmu 39](#_Toc297232493)

[6 Galvenie riski/iespējas 40](#_Toc297232494)

[7 plānotās sistēmas izveides izmaksas un to efektivitātes analīze 43](#_Toc297232495)

[7.1 Plānotās sistēmas izveides izmaksas 43](#_Toc297232496)

[7.2 Plānotie sistēmas izveides un uzturēšanas finansējuma avoti pa gadiem 43](#_Toc297232497)

[7.3 Alternatīvu analīze 43](#_Toc297232498)

[7.4 Projektu izmaksu efektivitātes analīze 44](#_Toc297232499)

# Par dokumentu

Šis dokuments ir izstrādāts, balstoties uz iepirkuma “ERAF projekta „Daudzvalodu korpusa un mašīntulkošanas infrastruktūras izveide e-pakalpojumu pieejamības nodrošināšanai” sistēmas koncepcijas izstrāde”, iepirkuma identifikācijas Nr. VAKIS 2010/1

## Izmantotie termini, akronīmi un saīsinājumi

1.tabula „Dokumentā izmantotie termini, akronīmi un saīsinājumi”

| **Termins, Saīsinājums** | **Skaidrojums** |
| --- | --- |
| API | Application Programming Interface |
| BLEU | Metrika automātiskai MT kvalitātes novērtēšanai[[4]](#footnote-5) <http://en.wikipedia.org/wiki/BLEU>  Metrika parāda cik mašīnas tulkojums ir līdzīgs cilvēka tulkojumam. Ja mašīna pārtulko 500 teikumus precīzi tieši tāpat kā cilvēks, tad BLEU metrika mašīnas tulkojumu novērtē ar 100 punktiem, ja mašīnas tulkojumā nav pat neviena kopīga vārda ar cilvēka tulkojumu, tad BLEU metrika šādu tulkojumu novērtē ar 0 punktiem. Savā ziņā var teikt, ka BLEU metrika parāda par cik procentiem mašīnas tulkojums sakrīt ar cilvēka tulkojumu. Parasti, ja BLEU metrika novērtē MT sistēmu ar 20 punktiem vai mazāk, tad šādu MT sistēmu var uzskatīt par samērā zemas kvalitātes sistēmu. Savukārt, ja MT sistēma ir novērtēta ar vairāk kā 40-50 BLEU punktiem, tad tā ir samērā labas kvalitātes MT sistēma. Reti kad kāda MT sistēma tiek novērtēta ar vairāk kā 50 BLEU punktiem. |
| CLARIN | CLARIN projekts (<http://www.clarin.lv>).  CLARIN ir plaša mēroga sadarbības iniciatīva, kurā piedalās daudzas Eiropas valstis. Tās uzdevums ir izveidot integrētu, sadarbību veicinošu pētniecības infrastruktūru, kas ļautu viegli piekļūt un izmantot valodas resursus un tehnoloģijas humanitāro, sociālo un eksakto zinātņu pētniekiem. CLARIN iniciatīvas sagatavošanas posms tiek īstenots ar 7. pamatprogrammas projekta CLARIN finansiālu atbalstu. Latvijas Universitātes Matemātikas un informātikas institūta dalību 7. pamatprogrammas projektā CLARIN finansiāli atbalsta LR Izglītības un Zinātnes ministrija. Latvijā plānotās aktivitātes sagatavošanās posmā saskaņojas ar kopējām CLARIN projekta aktivitātēm un tiek īstenotas kā atbilstošie nacionālie pasākumi CLARIN mērķu sasniegšanai. |
| EDS | Elektroniskās deklarēšanas sistēma |
| ERAF | Eiropas Reģionālās attīstības fonds |
| ES | Eiropas Savienība |
| IKT | Informācijas un komunikācijas tehnoloģijas |
| IVIS | Integrētā Valsts informācijas sistēma (skat. VISS) |
| LDC | Lauksaimniecības datu centrs |
| LAD | Lauku atbalsta dienests |
| LGPL | Lesser General Public License[[5]](#footnote-6) Atvērtā koda programmatūras licence, kas atļauj programmatūru brīvi izplatīt, ja vien tā netiek mainīta |
| LVRTC | Latvijas Valsts radio un televīzijas centrs |
| MT | Mašīntulkošana |
| MT logrīks | Veids kā MT serviss tiks integrēts interneta lapās (*MT widget*) |
| OCR | Optiskā rakstzīmju atpazīšana (*optical character recognition*) |
| RBMT | Likumos bāzētā mašīntulkošana (*rule based machine translation*) |
| SMT | Statistiskā mašīntulkošana |
| SOA | Servisorientēta arhitektūra (*Service Oriented Architecture*) |
| SOAP | Vienkāršs objektu pieejas protokols (*Simple Object Access Protocol)* |
| TTC | Tulkošanas un terminoloģijas centrs |
| VAAD | Valsts augu aizsardzības dienests |
| VID | Valsts ieņēmumu dienests |
| VISS | Valsts informācijas sistēmu savietotājs |
| VMD | Valsts meža dienests |
| VRAA | Valsts reģionālā attīstības aģentūra |
| VVC | Valsts valodas centrs |
| XML | Paplašināmās iezīmēšanas valoda (eXtensible Markup Language) |

# Esošās situācijas analīze

## Esošie procesi

Pašlaik, valsts pārvaldes iestādēs netiek izmantota vienota, esoša MT sistēma, atsevišķās iestādēs tiek izmantoti trešo pušu risinājumi, bet nav pieejami dati par šo sistēmu izplatību un izmantošanas biežumu.

Uz plānotās sistēmas darbību ir attiecināmi šādi esoši procesi:

1. valsts pārvaldes elektronisko pakalpojumu sniegšana – noteiktu valsts pārvaldes pakalpojumu sniegšana sabiedrībai, izmantojot [*www.latvija.lv*](http://www.latvija.lv) portālu. Kopumā dokumenta izstrādes laikā elektronisko pakalpojumu katalogā ir pieejami 26 pakalpojumi, šī procesa turētājs ir par konkrētā pakalpojuma sniegšanu atbildīgā iestāde, procesa mērķis ir atvieglot valsts iestāžu nodrošināto pakalpojumu pieprasīšanu un saņemšanu, nodrošinot to attālināti un izmantojot informācijas un komunikācijas tehnoloģiju sniegtās iespējas. Procesa rezultātā sabiedrība iegūst iespēju attālināti pieprasīt un saņemt dažādus valsts sniegtos pakalpojumus.
2. informācijas apmaiņa starp iestādēm un sabiedrību – saskaņā ar esošo normatīvo regulējumu, valsts pārvaldes pienākums ir nodrošināt, lai sabiedrībai būtu pieejama informācija, kura ir iestādes rīcībā vai kuru iestādei atbilstoši tās kompetencei ir pienākums radīt, kā arī veicināt sabiedrības līdzdalību valsts pārvaldē. Viens no šī procesa izpildes veidiem ir informācijas publicēšana elektroniskajos informācijas resursos (mājaslapas, blogi utt.), procesa mērķis ir nodrošināt informācijas apmaiņu starp valsts pārvaldes institūcijām un sabiedrību. Šī procesa turētājs ir katra iestāde atsevišķi savas kompetences ietvaros. Procesa rezultātā sabiedrība, izmantojot elektroniskos informācijas resursus, saņem tai nepieciešamo informāciju. Informācijas apmaiņas process ir cieši saistīts ar citiem valsts pārvaldes pamatprocesiem, jo ļauj operatīvi nodrošināt piekļuvi aktuālajai informācijai maksimāli plašam sabiedrības lokam.

Portāla [*www.latvija.lv*](http://www.latvija.lv) un tajā publicēto elektronisko pakalpojumu izmantojamība dokumenta izstrādes laikā ir:

* [www.latvija.lv](http://www.latvija.lv) 2010.gadā ir apmeklēts 558’682 reizes no 426’890 dažādiem datoriem
* [www.latvija.lv](http://www.latvija.lv) publicētie elektroniskie pakalpojumi 2010.gadā ir kopumā izmantoti 348’955 reizes

Informācija par elektroniskajiem pakalpojumiem un to apraksti ir pieejami tikai latviešu valodā.

Kopumā pašlaik Latvijā ir 16 augstākās valsts tiešās pārvaldes iestādes, kuru padotībā ir vairāk nekā 290 iestādes[[6]](#footnote-7) (pakļautībā esošas iestādes, akciju sabiedrības, pārraudzības iestādes utt.), kopumā tiek izmantotas vairāk nekā 100 mājaslapas, kurās galvenokārt tiek publicēta informācija valsts valodā. Informācija mājas lapās tiek aktualizēta latviešu valodā, atsevišķos gadījumos arī citās (angļu, krievu, franču), tomēr šāda tulkošana tiek veikta tikai ļoti nelielam informācijas apjomam un vairumā gadījumu, tā ir statistiska, kopējā informācija, kuru nav nepieciešams regulāri atjaunot, nevis aktualitātes. Tāpat ļoti reti ir novērojams dažādu pieteikumu formu un citu dokumentu, kas ir iegūstami no valsts pārvaldes mājas lapām, tulkojums dažādām iedzīvotāju grupām saprotamā valodā.

## Esošajos procesos iesaistītās personas

Procesā „Valsts pārvaldes elektronisko pakalpojumu sniegšana” ir iesaistītas šādas personas:

1. Esošais personāls – precīzs elektronisko pakalpojumu sniegšanā iesaistītā personāla skaits nav identificējams. Sistēmas kontekstā ir lietderīgi apskatīt IT personāla, kas ir tieši iesaistīts pakalpojumu uzturēšanā, skaitu, kas kopumā, koncepcijas apraksta veidošanas brīdī, nepārsniedz 40 cilvēkus. Šī personāla pienākumos ietilpst pakalpojuma tehniskā (nepieciešamās infrastruktūras, attiecīgo sistēmu u.c.) atbalsta nodrošināšana pakalpojuma sniedzēja iestādē.
2. Esošie klienti – privātpersonas, kas normatīvajos aktos noteiktajā kārtībā ir tiesīgas pieprasīt un saņemt elektroniskos pakalpojumus. Uz koncepcijas apraksta veidošanas brīdi par tādiem var tikt uzskatīta jebkura persona (arī personas, kas nav Latvijas iedzīvotāji), kurai ir interese izmantot kādu no publiskajiem e-pakalpojumiem, un visas personas, kurām ir iespēja autorizēties portālā [www.latvija.lv](http://www.latvija.lv) (koncepcijas veidošanas brīdī tie ir: E-ME (LVRTC nodrošinātā elektroniskā paraksta), *Swedbank*, *SEB*, Latvijas Krājbanka, Citadele banka, Norvik banka un *Lattelecom* mobilais ID lietotāji)
3. Ieinteresētās iestādes – pašlaik [www.latvija.lv](http://www.latvija.lv) reģistrētu elektronisko pakalpojumu sniegšanu nodrošina kopumā 9 iestādes (Pilsonības un migrācijas lietu pārvalde, Iekšlietu ministrijas Informācijas centrs, Valsts zemes dienests, Valsts policija, Tiesu administrācija, Veselības ekonomikas centrs, Veselības norēķinu centrs, Rīgas un Ventspils pilsētu pašvaldības), kā arī šo pakalpojumu sniegšanā ir iesaistīta Valsts reģionālās attīstības aģentūra. Tāpat elektroniskus pakalpojumus sniedz citas iestādes, kuras izmanto savas tehnoloģiskās iespējas elektronisko pakalpojumu nodrošināšanā un lietotāju pārvaldībā (piemēram: Valsts ieņēmumu dienests, Lauku atbalsta dienests, un citi).   
   Šo iestāžu līdzdalība esošajos procesos ir brīvprātīga, iestāžu pienākumus un tiesības nosaka katra konkrētā pakalpojuma darbību reglamentējošie normatīvie akti, kā arī vienošanās, kas tiek slēgta starp pakalpojuma sniedzēju un portāla [*www.latvija.lv*](http://www.latvija.lv) nodrošinātāju VRAA.

Procesā „Informācijas apmaiņa starp valsts pārvaldes iestādēm un sabiedrību” ir iesaistītas šādas personas:

1. Esošais personāls – sistēmas kontekstā par esošo personālu var tikt uzskatītas divas valsts pārvaldes iestāžu darbinieku grupas: IT personāls, kura pienākumos ietilpst IKT tehniskā atbalsta nodrošināšana iestādē, kā arī informācijas apmaiņā iesaistītais personāls (lietveži, komunikāciju nodaļu darbinieki, sabiedrisko attiecību speciālisti, u.c.), kuru pienākumos ietilpst iestādes informācijas publicēšana elektroniskajos informācijas līdzekļos. Ņemot vērā, ka iestādēs informācijas izvietošanas process ir organizēts dažādi, šī personāla precīzu skaitu nav iespējams noteikt.
2. Esošie klienti – sabiedrības locekļi (juridiskas un fiziskas personas, Latvijas un ārvalstu iedzīvotāji utt.), kas vēlas iegūt informāciju, kas ir pieejama valsts pārvaldes elektroniskajos informācijas resursos un kam uz to ir atbilstošas tiesības. Saskaņā ar spēkā esošo normatīvo regulējumu, valsts pārvaldes elektroniskajos informācijas resursos publicētā informācija tiek uzskatīta par neklasificētu informāciju un līdz ar to piekļuves tiesības šādai elektroniskai informācijai ir praktiski neierobežotam interesentu lokam. Vienīgais objektīvais ierobežojums ir attiecīgo IKT tehnoloģiju pieejamība.
3. Ieinteresētās iestādes – saskaņā ar spēkā esošo normatīvo regulējumu par ieinteresētajām iestādēm var uzskatīt praktiski visas valsts pārvaldes iestādes, kas informācijas atspoguļošanai un komunikācijai ar sabiedrību izmanto elektroniskos informācijas resursus. Ņemot vērā lielo ieinteresēto iestāžu skaitu nav iespējams aprakstīt visu iestāžu lomu esošajos procesos, tomēr kopumā šī procesa izpilde ir salīdzinoši standartizēta, iestādes izmanto to rīcībā esošos IKT infrastruktūru un resursus (serveri, datu pārraides iekārtas un līnijas, utt.) un publicē nepieciešamo informāciju attiecīgajā elektroniskajā informācijas resursā. Iestāžu pienākumi informācijas sniegšanā, kas izriet no normatīvā regulējuma prasībām, ir detalizēti aprakstīti nodaļā Nr.3.4 „Esošais tiesiskais regulējums”.

## Esošā sistēma

Šobrīd valsts pārvaldē netiek izmantota esoša MT sistēma, tulkošana notiek vai nu piesaistot attiecīgās jomas speciālistus, vai arī valsts pārvaldes darbinieki šo darbu veic pašu spēkiem.

### Esošie E-pakalpojumi un to tehniskie risinājumi

Šobrīd Latvijā, izmantojot portālu [www.latvija.lv](http://www.latvija.lv), ir pieejami 29 oficiāli publiski e-pakalpojumi. Esošie e-pakalpojumi ir būvēti, izmantojot Valsts informācijas sistēmu savietotāju (VISS), kurš sastāv no vairākām daļām. Uz MT attiecināmās daļas ir[[7]](#footnote-8):

* Publisko pakalpojumu reģistrs – informācija par pakalpojumiem, to sniegšanas veidiem
* E-pakalpojumu izpildes atbalsts (t.sk. katalogs) – reģistrs, kas satur e-pakalpojumus

Esošie tehniskie risinājumi nodrošina datu iegūšanu un reprezentāciju lietotājam standartizētā HTML formātā. Datu apstrāde ir veidota, izmantojot IVIS servisus, kas pamatā izmanto E-pakalpojumu standartā v.1.01 minētās XML datu struktūras, *SOAP* protokolu un *SOA* arhitektūru.

Šobrīd realizētais e-pakalpojumu mehānisms portālā [www.latvija.lv](http://www.latvija.lv) ir pieejams tikai latviešu valodā (tulkoti citās valodās nav ne pakalpojumu apraksti, ne arī sniedzamā informācija).

No izpētes gaitā identificētajiem, ārpus [www.latvija.lv](http://www.latvija.lv) sniedzamajiem un saņemamajiem e-pakalpojumiem vairums ir pieejami tikai latviešu valodā (piemēram: VID EDS sistēma, LAD E-Pieteikšanās sistēma, VAAD informācija par kaitēkļiem, VMD pakalpojumu katalogs un veidlapas, LR Ekonomikas ministrijas Būvkomersantu reģistrs, LDC publiskā datubāze un citi), un tikai daži (piemēram: ZVA zāļu reģistra daļa) ir daļēji pieejami citās valodās.

## Esošais tiesiskais regulējums

Esošos procesus regulē šādi nacionālā līmeņa normatīvie akti:

1. Informācijas atklātības likums[[8]](#footnote-9) - likuma mērķis ir nodrošināt, lai sabiedrībai būtu pieejama informācija, kura ir iestādes rīcībā vai kuru iestādei atbilstoši tās kompetencei ir pienākums radīt. Šis likums nosaka vienotu kārtību, kādā privātpersonas ir tiesīgas iegūt informāciju iestādē un to izmantot;
2. 2007.gada 6.marta MK noteikumi Nr.171 „Kārtība, kādā iestādes ievieto informāciju internetā”[[9]](#footnote-10) – reglamentē kādā iestādes ievieto informāciju internetā, lai nodrošinātu tās pieejamību, tajā skaitā minimālo mājaslapas sadaļu skaitu un tajās publicējamo informāciju, kā arī informācijas apjomu, kas ir jāsniedz svešvalodās;
3. Iesniegumu likums[[10]](#footnote-11) - likuma mērķis ir veicināt privātpersonas līdzdalību valsts pārvaldē;
4. Valsts valodas likums[[11]](#footnote-12) – likums reglamentē latviešu valodas lietošanu privātpersonu un valsts pārvaldes iestāžu oficiālajā komunikācijā. Plānotās sistēmas darbībai ir jāievēro likumā noteiktās normas un jānodrošina tikai esošo procesu atsevišķu posmu (saskarne ar lietotājiem) realizācija tiem saprotamākā vai ērtākā veidā (valodā).

Kā arī šādi ES līmeņa normatīvie akti:

Eiropas Savienības Padomes Rezolūcija par Eiropas multilingvisma stratēģiju[[12]](#footnote-13), kas nosaka Eiropas valodu daudzveidības nodrošināšanas uzdevumus, tai skaitā aicina dalībvalstis attīstīt un izmantot tulkošanas tehnoloģijas, lai attīstītu mobilitāti Eiropas Savienībā un informācijas un zināšanu izplatību.

## Esošie politikas dokumenti

Esošos procesus daļēji ietekmē šādi nacionālā līmeņa politikas plānošanas dokumenti:

1. Valsts programma „Sabiedrības integrācija Latvijā”[[13]](#footnote-14), kas nosaka, ka Latvijas valsts visu iedzīvotāju līdzdalība ir būtisks solis demokrātiskas, tiesiskas valsts pastāvēšanai un Integrācijas mērķis ir veidot demokrātisku, saliedētu pilsonisku sabiedrību, kas balstās uz kopīgām pamatvērtībām.

Esošos procesus ietekmē šādi ES līmeņa politikas plānošanas dokumenti:

1. Eiropas Komisijas paziņojums „Daudzvalodība: Eiropas priekšrocība un kopīga apņemšanās”[[14]](#footnote-15), kas nosaka, ka Eiropas Savienībai ir jāpielāgojas pārmaiņām mūsdienu globalizētajā pasaulē, pievēršoties iespēju, pieejamības un solidaritātes pamatprincipiem. Daudzvalodīgā Eiropas Savienībā tas nozīmē, ka:
   1. ikvienam nepieciešama iespēja pienācīgi sazināties, lai īstenotu savu potenciālu un vislabākajā veidā izmantotu modernas un inovatīvas Eiropas Savienības piedāvātās iespējas;
   2. ikvienam jābūt pieejamai atbilstošai valodu apmācībai vai citiem līdzekļiem, kas palīdz sazināties, lai valoda neradītu nevajadzīgus šķēršļus dzīvošanai, strādāšanai vai saziņai Eiropas Savienībā;
   3. ievērojot solidaritātes principu, arī tiem, kuri nevar apgūt citas valodas, nodrošināmi atbilstoši saziņas līdzekļi, kas darītu pieejamu daudzvalodu vidi;
2. Pasaules Informācijas sabiedrības samita (*WSIS* - *World Summit on the Information Society*) deklarācija un rīcības plāns, kas deklarē pasaules valstu un globālo organizāciju apņemšanos izveidot informācijas sabiedrību, kas balstās uz vispārēju pieeju informācijai un vienlīdzīgu iespēju sniegšanu visām sabiedrības grupām (t.sk. lokālo valodu lietojumu IKT pakalpojumos).

## Mūsdienīgu mašīntulkošanas risinājumu analīze

### MT attīstības vēsture

Mašīntulkošanai ir bagāta vēsture, kas sākusies pagājušā gadsimta 40. gados. Izmantot datorus dabīgās valodas tekstu tulkošanā pirmo reizi ierosināja Vorens Vīvers (*Warren Weaver*) 1947. gadā, bet pirmos mašīntulkošanas demonstrējumus (sešdesmit teikumi tika automātiski iztulkoti no krievu valodas angļu valodā) plašākai publikai 50. gadu sākumā veica Džordžtaunas Universitāte un *IBM*[[15]](#footnote-16). Kopš tā laika ir veikti daudzi eksperimenti un radītas daudzas eksperimentālas un praktiskas sistēmas darbam ar dažādām valodām un valodu pāriem.

Neraugoties uz ilgo mašīntulkošanas vēsturi, tās sākotnējais mērķis — tulkotāju–cilvēku aizstāšana — joprojām nav sasniegts: pašreizējās sistēmas vēl ir tālu no tā, lai radītu tāda paša līmeņa tulkojumus, kā tulkotāji–cilvēki[[16]](#footnote-17). Tomēr šodien mašīntulkošana ir perspektīva datorlingvistikas joma ar savu teoriju un praksi un aptuveni 200 mašīntulkošanas sistēmām un lietojumiem[[17]](#footnote-18). Mašīntulkošanas attīstību virza tās lielais zinātniskais, politiskais, ekonomiskais, sociālais un kultūrpolitiskais nozīmīgums. Pēdējā desmitgadē mašīntulkošana ir bijis īpaši strauji augošs un perspektīvs informācijas tehnoloģiju virziens. Jau šodien mašīntulkošanas sistēmas plaši izmanto gan individuāli lietotāji, gan dažādu nozaru speciālisti, lai uztvertu teksta jēgu svešvalodā, kā arī profesionāli tulkotāji, lai tulkošanas procesu padarītu efektīvāku un produktīvāku.

### Esošie MT risinājumi

Pieprasījums pēc ātras un lētas tulkošanas ir devis impulsu radīt vairākus komerciālus produktus (piemēram, *Systran, Promt, Reverso, LanguageWeaver*) un vairākus tulkošanas risinājumus, kas ir brīvi pieejami kā pakalpojumi tīmeklī un kam ir pieņemama tulkojumu kvalitāte plaši lietotajās valodās (vairākas kompānijas piedāvā MT servisus daudzām valodām, tai skaitā arī latviešu valodai, populārākie no tiem ir *Google Translator*[[18]](#footnote-19), *Microsoft Translator*[[19]](#footnote-20) un Tildes Tulkotājs[[20]](#footnote-21)). Šie risinājumi ir atvērtas un universālas sistēmas, kas lielākoties piedāvā tikai universālu tulkošanas rīku un neder dažādiem specifiskiem lietojumiem, piem., konkrētas nozares dokumentu tulkošanai. Parasti šos tiešsaistes tulkošanas risinājumus izmanto gadījuma rakstura lietotāji īsu tekstu tulkošanai. Gatavā veidā šos servisus nav iespējams pielāgot e‑pakalpojumu jomai. E‑pakalpojumu pieejamības nodrošināšanai šādas vispārīga lietojuma MT sistēmas nav izmantojamas, jo nenodrošina e‑pakalpojumu un valsts pārvaldes specifikai nepieciešamās tulkošanas kvalitātes, tehniskās un drošības prasības. Lai iegūtu kvalitatīvāku tulkojumu, MT ir jāpielāgo konkrētā lietojuma gramatiskajām, terminoloģiskajām un stilistiskajām prasībām.

Ļoti svarīgs aspekts MT e‑pakalpojumu jomā ir drošība. Ar e‑pakalpojumiem iedzīvotāji un juridiskas personas var piekļūt privātai, nepubliskai vai pat slepenai informācijai. No drošības viedokļa nav pieļaujams šādu informāciju tulkot, izmantojot publiskus bezmaksas servisus. Tie uzkrāj tulkoto informāciju un nenodrošina personas datu un konfidenciālas informācijas aizsardzību. Tādējādi ir jāizvēlas atbilstoša stratēģija šādas specifiskas mašīntulkošanas sistēmas izstrādei, kas nodrošina nepieciešamo tulkošanas kvalitāti un apmierina datu drošības, informācijas aizsardzības un sistēmas veiktspējas prasības.

### Mašīntulkošanā izmantotās metodes

Mašīntulkošana ir daudzdisciplināra, un ietver tādas jomas kā tulkošanas teorija, mākslīgais intelekts, komunikāciju teorija, lingvistika, matemātika u.c. Mašīntulkošanas sistēmu izstrādē tiek izmantotas dažādas metodes[[21]](#footnote-22). Esošās MT sistēmas var iedalīt pēc izmantotās tulkošanas stratēģijas— lingvistiskās (balstās uz likumiem/zināšanām) vai empīriskās (balstās uz piemēriem un tulkojumu statistiskās analīzes)[[22]](#footnote-23).

Lingvistiskā stratēģija jeb uz likumiem balstītā mašīntulkošana (*RBMT* — *rule-based machine translation*) ietver dažādas metodes, kuru pamatā ir lingvistiskas un ekstralingvistiskas zināšanas. Trīs galvenās metodes ir tiešais tulkojums, pārnesē balstītā tulkošana (*transfer-based MT*) un tulkošana caur starpvalodu (*interlingual MT*). Ar pārnesē balstītās tulkošanas metodi ir iegūti atzīstami rezultāti daudzās komerciālajās sistēmās. *RBMT* izmanto cilvēka radītus likumus un vārdnīcas, lai aprakstītu gan dabīgo valodu gramatiku, gan tulkošanas likumus. Tās spēj labi tulkot sintaktiski un leksiski vienkāršus tekstus ar viennozīmīgu semantiku, kuru tulkošanai tajās formalizētas visas nepieciešamās zināšanas. Taču tās slikti tulko plašāka spektra tekstus un to izstrāde un attīstīšana ir ļoti dārga.

Ekonomiskais izdevīgums ir viens no galvenajiem iemesliem, kādēļ mašīntulkošanā par dominējošo ir kļuvusi empīriskā paradigma. Tā ir sevi pierādījusi kā visefektīvākais risinājums gan laika un darbietilpības, gan tulkojuma kvalitātes ziņā. Statistisko mašīntulkošanu (*SMT*) pirmajā vietā izvirzīja jau paši pirmie rezultāti, kas tika iegūti 20. gs. 80. gadu beigās IBM projekta *Candide* pētījumos par angļu-franču *SMT* sistēmu[[23]](#footnote-24). Pēdējās desmitgades laikā daudzi pētījumi ir pievērsušies tieši *SMT*[[24]](#footnote-25).

*SMT* sākumā modeļu pamatā bija vārdi, bet, ieviešot frāžu modeļus[[25]](#footnote-26), tika sperts nozīmīgs solis uz priekšu. Jaunākajos pētījumos iestrādātas arī sintaktiskas un kvazisintaktiskas struktūras[[26]](#footnote-27). Viskvalitatīvākās *SMT* sistēmas (piemēram, *ATS*[[27]](#footnote-28), *CMU*[[28]](#footnote-29) un *IBM*[[29]](#footnote-30) sistēmas) izmanto frāžu tulkošanas metodes. Jaunākās SMT tendences atspoguļo darbs pie lingvistiskas anotācijas iestrādāšanas vārdu līmenī ar faktorizētiem tulkošanas modeļiem[[30]](#footnote-31) vai koka veida modeļiem[[31]](#footnote-32). Tas uzlabo mašīntulkošanas kvalitāti īpaši tajās valodās, kam ir bagātīga morfoloģija un brīva vārdu kārtība, kā arī palīdz risināt tādas problēmas kā atstatu vārdu secības pārkārtošana un teikumu līmeņa gramatiskā saskaņotība.

Ar statistikas metodēm *SMT* analizē iepriekš tulkotus tekstus (tā saucamā *trenēšana*), izveido statistiskos modeļus un tos izmanto jaunu tekstu tulkošanai. *SMT* izveidošana un pielāgošana ir daudz lētāka par *RBMT*. Lai izveidotu vienkāršu SMT sistēmu, ir nepieciešamas platformas sistēmu trenēšanai un darbināšanai un liels tekstu korpuss sistēmu trenēšanai. Akadēmiskajā vidē izveidotas labas atvērtā koda platformas SMT trenēšanai un darbināšanai – *Moses* MT rīkkopa, *Giza++* tekstu sastatīšanas rīkkopa, *SRILM* valodas modelēšanas rīkkopa[[32]](#footnote-33) un citi atvērti risinājumi. Piemēram, *Moses* MT rīkkopa ir pilnīga tulkošanas sistēma, kas tiek izplatīta ar *LGPL* (*Lesser General Public License*) licenci. Tajā ir visi nepieciešamie komponenti, lai veiktu datu priekšapstrādi un trenētu tulkošanas un valodu modeļus. *Moses* MT rīkkopai ir aktīva pētnieku kopiena; un kopš 2007. gada 1. marta tā ir lejupielādēta vairāk nekā 1000 reižu.

ES 7.ietvarprogrammas projekts *EuroMatrixPlus*[[33]](#footnote-34) ir parādījis, kā atvērtā koda rīkus un publiski pieejamos datus var izmantot, lai izveidotu SMT sistēmas visu ES oficiālo valodu pāriem. Atvērtā koda platformas *SMT* trenēšanai un darbināšanai ir izmantojamas arī latviešu valodai, trenējot uz ļoti liela apjoma latviešu, angļu un krievu paralēlā teksta korpusiem. Tulkošanas kvalitātes uzlabošanai šīm platformām jāizstrādā valodu specifiski moduļi. Tātad MT e-pakalpojuma vajadzībām nepieciešamo sistēmu var veidot, izmantojot esošās atvērtā koda platformas, bet, lai sistēma būtu pielāgota konkrētajai jomai un tulkošanas virzienam, ir nepieciešams liels tekstu korpuss un *SMT* sistēma ir īpaši jāpielāgo tulkošanas virzienam, to papildinot ar zināšanām par valodu – morfoloģiju, teikuma sintaksi u.tml.

# Izmaiņu pamatojums un būtība

Koncepcijas veidošanas laikā visa informācija, kas ir pieejamapotenciālajiem elektronisko pakalpojumu lietotājiem, kā arī lielākā daļa no valsts pārvaldes iestāžu informatīvajos resursos publicējamās aktuālās informācijas ir pieejama tikai latviešu valodā. Ieviešot MT sistēmu, tiks nodrošināta:

1. automātiska esošo un plānoto e-pakalpojumu informācijas tulkošana, padarot šos pakalpojumus pieejamākus plašākai sabiedrības daļai, tajā pašā laikā ievērojot normatīvo aktu prasības attiecībā uz valsts valodas lietošanu;
2. risinājums, kas ļaus valsts pārvaldes mājas lapu apmeklētājiem veikt tajās esošās informācijas kvalitatīvu tulkošanu uz angļu un krievu valodām.

Projekta mērķus var definēt šādi:

1. samazināt klātienē sniedzamo publisko pakalpojumu apjomu, nodrošinot lielāku e-pakalpojumu izmantošanu;
2. sekmēt e-pakalpojumu izplatību, padarot informāciju pieejamu plašākai sabiedrības daļai;
3. veicināt sabiedrības informētību par valsts pārvaldes procesiem un veicināt tās līdzdalību, nodrošinot informācijas pieejamību dažādām sabiedrības grupām;
4. veicināt sabiedrības iespēju lasīt informāciju citās valodās, izmantojot MT sistēmu.

Projekta rezultātā izveidotās sistēmas galvenie lietotāji būs valsts pārvaldes iestādes un to klienti – ministrijas un pašvaldības, kā arī to padotības iestādes un kapitālsabiedrības, uz kurām ir attiecināms esošais normatīvais regulējums par informācijas sniegšanu iedzīvotājiem.

Kā papildus ieguvumu var minēt esošo kultūrvēsturisko vērtību pieejamības paplašināšanu, jo dažādiem interesentiem būs iespēja iepazīties ar dažādiem materiāliem tiem saprotamā valodā, piemēram – Latvijas Nacionālās bibliotēkas digitalizētajiem materiāliem (padomju laika periodika, latviešu oriģinālliteratūra u.c. materiāli).

Projekta rezultātus varēs izmantot latviešu valodas tehnoloģiskās attīstības nodrošināšanā, valodas resursu bāzes izveidē tālākām akadēmiskām un lietišķām izstrādēm un integrācijai kopējā ES valodu resursu infrastruktūrā.

Lai nodrošinātu sistēmas pilnvērtīgi izmantošanu un turpmākās attīstības iespēju, ir nepieciešams veikt arī attiecīgas izmaiņas vairākos normatīvos aktos (informācija par nepieciešamajām izmaiņām ir pieejama sadaļā Nr. ).

# Plānotās situācijas apraksts

## Plānotie procesi

### Analīze

Plānotie MT procesi ir iedalāmi:

* Plānotie procesi, lai nodrošinātu MT iedarbināšanu;
* Plānotie procesi MT sistēmas darbināšanā.

Lai izveidotu MT servisu integrēšanai e‑pakalpojumos, ir nepieciešami procesi:

2.tabula „MT izveidošanai plānotie procesi”

|  |  |  |  |
| --- | --- | --- | --- |
| **Nr.** | **Process** | **Mērķi** | **Rezultāti** |
|  | Valodu korpusa izveide | Izveidot mašīntulkošanas sistēmas prasībām un projekta specifikai atbilstošu monolingvālo un paralēlo valodu korpusu (latviešu-angļu, latviešu-krievu) | Pietiekams vārdu un teikumu kopums MT sistēmas attīstīšanai |
|  | jāizveido bāzes MT serviss, izmantojot pieejamās SMT platformas un jau pieejamos paralēlo un monolingvālo tekstu korpusus; | Izveidot atbilstošo programmnodrošinājumu un notestēt tā darbību | Izstrādāta un notestēta programmatūra |
|  | jāpielāgo SMT sistēmas konkrētiem valodu pāriem un e-pakalpojumu specifikai; | Pielāgot izveidoto MT servisu nepieciešamajai valodas specifikai | Izveidotais risinājums spēj nodrošināt kvalitatīvu, plānotajiem teksta korpusiem piemērotu tulkošanu. |
|  | jāizveido MT servisam nepieciešamā infrastruktūra (gan programmatūras, gan aparatūras); | Panākt izvirzītajiem kvalitātes kritērijiem (piem. ātrumiem, pieejamībai) atbilstošu MT servisa darbību | Iegādāta, uzstādīta (iedarbināta) un notestēta MT servisam nepieciešamā infrastruktūra |
|  | MT serviss jāieintegrē e-pakalpojumu vidē. | Uzlabot e-pakalpojumu vidi un palaist darbībā MT servisu | MT e-pakalpojums ir integrēts esošo e-pakalpojumu vidē gan kā autonoms serviss, gan integrēts esošo/plānoto e-pakalpojumu infrastruktūrā |

Lai nodrošinātu MT e-pakalpojuma ilgtspēju, ir nepieciešami šādi procesi:

3.tabula „MT darbināšanai plānotie procesi”

|  |  |  |  |
| --- | --- | --- | --- |
| **Nr.** | **Process** | **Mērķi** | **Rezultāti** |
|  | MT iekļaušana jaunos e-pakalpojumos un plašs izveidotā MT logrīka lietojums | Paplašināt MT lietojumu e‑pakalpojumos | Visos jaunajos e-pakalpojumos tiek izmantots MT serviss tulkošanas nodrošināšanai.  MT logrīks tiek izmantots arī citos valsts pārvaldes publiskotajos materiālos |
|  | Valodu korpusa papildināšanas process | Nodrošināt tulkojumu kvalitātes pastāvīgu uzlabošanu un MT pielaisto kļūdu samazināšanos | MT tulkojuma kvalitāte arvien vairāk pietuvojas cilvēka veiktam kvalitatīvam tulkojumam |
|  | Infrastruktūras un programmatūras uzturēšanas process | Nodrošināt MT servisa darbību | MT serviss ir pieejams |

### Esošo valodas korpusu izpēte

Lai izveidotu angļu-latviešu MT sistēmu, ir nepieciešams gan angļu-latviešu paralēlais korpuss, gan latviešu monolingvālais korpuss. SMT platformas izmanto paralēlo korpusu, lai no tā izgūtu vārdu un frāžu tulkojumus un to varbūtības, un monolingvālo korpusu, lai no tā izgūtu informāciju par izejas valodas tekstu, piemēram, vārdu un frāžu secību, vārdu savstarpējo saskaņojumu u.tml. Lai izveidotu sākuma līmeņa MT sistēmu šaurā jomā, ir nepieciešami vismaz 1,5 miljoni teikumu (apmēram 30 miljonu vārdu) dotās jomas paralēlais korpuss un vismaz 15 miljonu teikumu monolingvālais korpuss. MT sistēmu izveidē ir svarīgi gan vispārīga lietojuma korpusi, gan tieši jomai specifiski korpusi.

Tiešā veidā mašīntulkošanā var izmantot tikai tādu paralēlu korpusu, kas ir apkopots un sastatīts teikumu līmenī. Vienkārši dokumenti un to tulkojumi otrā valodā nav tiešā veidā izmantojami, tie vispirms ir jāapstrādā - no tiem ir jāizvelk teksts un jāsastata pa teikumiem.

#### Angļu-latviešu paralēlais korpuss

Latvijā līdz šim nav veiktas mērķtiecīgas un valsts atbalstītas aktivitātes angļu-latviešu paralēlā korpusa izveidē. Ir vairākas aktivitātes kas ir notikušas Eiropas kontekstā un kuru rezultātā ir izveidoti arī vairāki angļu-latviešu paralēlie korpusi.

Apjomīgākais un zināmākais no šiem korpusiem ir *JRC-ACQUIS* daudzvalodu korpuss[[34]](#footnote-35), kurā kā viena no korpusa 22 valodām ir arī latviešu valoda. Šajā korpusā ir apkopoti ES dokumentu tulkojumi, kas ir veikti Latvijai iestājoties ES. Korpusā ir apmēram 1,3 miljoni paralēlu angļu latviešu teikumu.

Cits ar ES dokumentāciju saistīts daudzvalodu korpuss, kurā ir arī latviešu valoda, ir *DGT-TM* daudzvalodu korpuss[[35]](#footnote-36). Arī šajā korpusā ir ES dokumentu tulkojumi. Korpusā ir apmēram 1,1 miljons paralēlu angļu-latviešu teikumu. Šis korpuss lielā mērā pārklājas ar *JRC-ACQUIS* korpusu, taču tas ir izveidots citādi un tajā ir teksti, kas nav iekļauti *JRC-ACQUIS*.

Angļu-latviešu paralēlais korpuss ir pieejams arī *OPUS* daudzvalodu korpusā[[36]](#footnote-37). *OPUS* korpusā ir apkopoti paralēlie teksti no dažādām jomām, tai skaitā, programmatūras lokalizācija, filmu subtitri, medicīna u.c. Latviešu valoda vērā ņemamā apjomā ir pārstāvēta tikai *OPUS* ietilpstošajā *EMEA* korpusā[[37]](#footnote-38), kas satur apmēram 300 tūkstošus paralēlus angļu-latviešu teikumus medicīnas jomā.

Minētie paralēlie angļu-latviešu korpusi jau apkopoti, sastatīti teikumu līmenī, ir derīgi izmantošanai MT sistēmu izveidošanai un ir publiski pieejami. Eksistē arī citi paralēli teksti, kas vai nu nav apkopoti un sastatīti, vai nav publiski pieejami. Piemēram, ne visi teksti, kas ir tulkoti ES institūcijās, ir iekļauti *DGT-TM* un *JRC-ACQUIS* korpusos, ir arī daudz ar Latvijas likumdošanu un valsts pārvaldi saistītu tekstu tulkojumi, ko veic VVC[[38]](#footnote-39) (iepriekš TTC). Lielākais paralēlo tekstu avots var būt grāmatas, kas ir tulkotas no angļu valodas latviski un otrādi. Ļoti daudz teksta tiek tulkots dažādās firmās, kas sniedz tulkošanas un lokalizācijas pakalpojumus.

Tildes vadītajā ES 7.ietvarprogrammas projektā *ACCURAT* tiek izstrādātas metodes paralēlu teikumu un frāžu izgūšanai no tīmekļa[[39]](#footnote-40). Tilde sadarbībā ar *Microsoft Research* izveidojusi liela apjoma latviešu-angļu tīmekļa tekstu korpusu, ko šie uzņēmumi izmanto mašīntulkošanas tehnoloģiju pētniecībā[[40]](#footnote-41). Mašīntulkošanas izstrādē izmantojamas arī tulkojošās un terminoloģiskās vārdnīcas, piemēram, *EuroTermBank*[[41]](#footnote-42).

Lai palīdzētu automatizēt tulkošanas procesu, virkne globālu un reģionālu IT uzņēmumu (*Adobe, Oracle, Sun, Intel*, Tilde, *Microsoft* u.c.) ir izveidojuši organizāciju *TAUS Data Associaton*[[42]](#footnote-43) (*TAUS DA*). Tā veido datubāzi ar šo un citu uzņēmumu produktu tekstu tulkojumiem dažādās valodās. *TAUS DA* datubāzē ir apkopoti tulkojumi, kas nāk no dažādām organizācijām - gan uzņēmumiem, gan ES institūcijām, gan individuāliem tulkotājiem. Šie tulkojumi ļoti noder gan tulkotāju darba efektivitātes celšanai, gan mašīntulkošanas sistēmu uzlabošanai. *TAUS DA* datubāzē ir arī daudz angļu-latviešu paralēlo tekstu, tai skaitā tādu, kas nav pieejami citos avotos. *TAUS DA* datubāze ir pieejama tikai *TAUS DA* biedriem.

#### Krievu-latviešu paralēlais korpuss

Krievu-latviešu valodu pāris vērā ņemamā apjomā nav pārstāvēts nevienā no iepriekš minētajiem daudzvalodu paralēlajiem korpusiem, un nav arī zināmi citi paralēlie korpusi, kuros šis valodu pāris būtu pārstāvēts.

Taču eksistē liels apjoms neapkopotu un nesastatītu krievu-latviešu tulkoto tekstu. Šādu tekstu piemēri ir Latvijas krievu kopienas veidoti un tulkoti teksti, ziņas, kas informē par notikumiem Latvijā gan krievu, gan latviešu valodā, padomju laika teksti, ļoti daudz teksta tiek tulkots dažādos uzņēmumos, kas sniedz tulkošanas un lokalizācijas pakalpojumus. Lielākais avots krievu-latviešu paralēlā korpusa izveidei ir grāmatas, kas ir tulkotas no krievu valodas latviski un otrādi.

#### Latviešu monolingvālais korpuss

Latviešu valodas korpusa izveide ietverta starp galvenajiem valsts valodas politikas īstenošanas uzdevumiem (Ministru kabineta 2010.g. 11.augusta rīkojums Nr.470 par grozījumiem Ministru kabineta 2005.g. 2.marta rīkojumā Nr.137 "Par Valsts valodas politikas pamatnostādnēm 2005.–2014.gadam"). 2005. gadā Latvijas Universitātes Matemātikas un Informātikas Institūtā (LU MII) ar Valsts Valodas centra atbalstu tika izstrādāta „Latviešu valodas korpusa koncepcija”. Koncepcija apraksta korpusa veidošanas pamatprincipus.

Atbilstoši koncepcijai ir izveidots „Līdzsvarots mūsdienu latviešu valodas tekstu korpuss”, kura pašreizējā versija satur apmēram 3,5 miljoni vārdlietojumu. Korpusā iekļauti drukātie un elektroniskie materiāli, kas radušie pēc 1991. gada. Nesen korpuss papildināts ar automātisku morfoloģisko marķējumu. Korpuss izmantojams zinātniskiem mērķiem, lietojot programmu *Bonito*, kas ļauj korpusā meklēt, bet neļauj to lejuplādēt un izmantot SMT trenēšanai.

LU MII tāpat izveidojis Latviešu valodas tīmekļa korpusu (aptuveni 100 milj. vārdlietojumu) un Latvijas Republikas 5.-9. Saeimas sēžu stenogrammu korpusu (vairāk nekā 20 milj. vārdlietojumu). Tāpat kā līdzsvarotais korpuss, arī šie korpusi ir izmantojami caur *Bonito* pārlūkprogrammu.

Līdztekus iepriekš minētajiem korpusiem LU MII ir uzkrājis apjomīgas tekstu kolekcijas latviešu valodā. Kā nozīmīgākās minamas: Latviešu literatūras klasika, Latviešu literatūras zelta fonds un Folkloristikas elektroniskā bibliotēka. Pilns LU MII resursu uzskaitījums dots <http://valoda.ailab.lv>.

Latviešu valodas korpusa satura izveidē ir iesaistīta arī Latvijas Nacionālā bibliotēka, kā piemēru var minēt periodikas sadaļu Nacionālajā digitālajā bibliotēkā. Taču arī šajā korpusā ir iespējams tikai meklēt un tas tiešā veidā bez papildus apstrādes nav izmantojams SMT trenēšanai.

Liela apjoma latviešu daiļliteratūras un enciklopēdiskās literatūras korpusu izveidojusi Tilde. Tas pieejams individuālu vienumu lasīšanas režīmā portālā <http://letonika.lv> .

Arī citās Latvijas zinātniskajās un mācību iestādēs tiek uzkrāti dažādi latviešu valodas teksti. Pirms diviem gadiem *CLARIN* projektā[[43]](#footnote-44) apzinātie valodas resursi atrodami: <http://valoda.ailab.lv/clarin/resursu_un_riku_parskats.jsp>.

#### Angļu monolingvālais korpuss

Angļu valodas korpusi ir daudz pieejamāki, eksistē vairāki apjomīgi, plaši zināmi un pieejami angļu valodas korpusi, piemēram, *American National Corpus*[[44]](#footnote-45) (22 milj. vārdu), *Bank of English*[[45]](#footnote-46) (525 milj. vārdu), *British National Corpus*[[46]](#footnote-47) (100 milj. vārdu), *Corpus of Contemporary American English*[[47]](#footnote-48) (410 milj. vārdu), *International Corpus of English*[[48]](#footnote-49), *Oxford English Corpus*[[49]](#footnote-50) (vairāk kā 2 mljrd. vārdu), *Leipzig Corpora Collection* (21 milj. vārdu) un daudzi citi. SMT trenēšanā plaši tiek izmantoti arī ziņu korpusi, piemēram*, News Commentary Corpus*[[50]](#footnote-51) (440 milj. vārdi) un *English Gigaword corpus*[[51]](#footnote-52) (1,6 mljrd. vārdu), *Europarl* daudzvalodu korpusa[[52]](#footnote-53) angļu daļa (51 milj. vārdu) un automātiski no interneta savākti angļu tekstu korpusi, piemēram, *Google* n-gramu korpuss[[53]](#footnote-54) (apmēram 1 trilj. vārdu).

Trenējot jomai pielāgotu SMT sistēmu, ir svarīgi izmantot arī jomai specifisku monolingvālo korpusu. ePakalpojumu jomai vistuvākais pieejamais angļu monolingvālais korpuss ir *Europarl* korpusa angļu daļa.

#### Krievu monolingvālais korpuss

**Krievu valodas nacionālais korpuss[[54]](#footnote-55)** satur aptuveni 140 miljonu vārdlietojumus un ir paredzēts galvenokārt zinātniskiem pētījumiem krievu valodas leksikā, gramatikā un vēsturē, kā arī mācību mērķiem. Korpuss ir reprezentatīvs (satur dažāda tipa runas un teksta resursus dažādos žanros, ieskaitot daiļliteratūru, publicistiku, mācību, zinātņu, lietišķu, kā arī dialektu tekstus); sabalansēts (resursu laika un skaitļa ziņā); anotēts (satur metateksta, morfoloģisku, akcentu un semantisku anotāciju). Esošo korpusu ir plānots paplašināt līdz 200 miljonu vārdiem un papildināt ar sintaksisku anotāciju[[55]](#footnote-56). Korpuss ir pieejams meklēšanai internetā caur korpusa meklēšanas lapu[[56]](#footnote-57) un nav paredzēts izplatīšanai. Pašlaik korpuss nav pieejams lejupielādēšanai.

**Krievu literārās valodas korpuss[[57]](#footnote-58)** [[58]](#footnote-59)satur aptuveni miljonu vārdlietojumus un ir paredzēts galvenokārt zinātniskiem pētījumiem. Korpuss ir reprezentatīvs (satur dažāda tipa teksta resursus dažādos žanros, ieskaitot daiļliteratūru, publicistiku un zinātņu tekstus); sabalansēts (resursu laika un skaitļa ziņā); anotēts (satur morfoloģisku anotāciju). Korpuss nav pieejams lejupielādēšanai.

**Krievu tekstu Upsalas korpuss[[59]](#footnote-60)** ietver 600 daiļliteratūras un informatīvus tekstus krievu valodā (korpusa apjoms ir miljons vārdlietojumu). Informatīvie teksti ir dažādās nozarēs no 1985. gada līdz 1989. gadam (ekonomika, politika, sociālās zinības, vēsture, likumdošana u.c.). Daiļliteratūras korpusa daļa ietver aptuveni 40 autoru darbus. Korpuss nav anotēts. Korpuss nav pieejams lejupielādēšanai.

**Krievu valodas avīžu korpuss[[60]](#footnote-61)** satur aptuveni miljonu vārdlietojumus no avīzēm krievu valodā 20. gadsimta beigās. Korpuss ir anotēts (žanru anotācija). Korpuss ir pieejams meklēšanai internetā caur korpusa meklēšanas lapu[[61]](#footnote-62). Korpuss nav pieejams lejupielādēšanai.

**Krievu valodas Tībingenes korpuss[[62]](#footnote-63)** satur aptuveni 25 miljonus vārdlietojumus. Korpusa resursi: Upsalas korpuss, periodika, daiļliteratūra. Korpuss ir automatizēti anotēts ar morfoloģisku[[63]](#footnote-64). Korpuss ir paredzēts meklēšanai internetā. Korpuss nav pieejams lejupielādēšanai.

**Krievu valodas uzziņas korpuss[[64]](#footnote-65) [[65]](#footnote-66)** (*Russian Reference Corpus*): pilota versija ietvēra aptuveni 35 miljonus vārdlietojumus. 45% no korpusa resursiem ir daiļliteratūra, un pārējie ir dažādu nozaru teksti (ieskaitot avīžu, zinātņu un citu žanru tekstus). Korpuss ir lematizēts un morfoloģiski anotēts. Korpuss ir potenciāli pieejams lejupielādēšanai.

**Krievu valodas interneta korpuss[[66]](#footnote-67) [[67]](#footnote-68)**. Korpusa avoti ir interneta resursi. Korpuss ir potenciāli pieejams lejupielādēšanai.

#### Apkopojums par esošajiem valodas korpusiem

Apkopojot iepriekš rakstīto, var secināt, ka:

* jau eksistē gatavi korpusi, kas nepieciešami, lai izveidotu sākuma līmeņa angļu-latviešu un latviešu-angļu SMT sistēmas, tās gan nebūs labi piemērotas valsts pārvaldes specifikai,
* situācija ir sliktāka ar latviešu-krievu virzienu, pašlaik vēl nav gatavu korpusu, ko varētu izmantot pat sākuma līmeņa SMT izveidei, rezultātā latviešu-krievu korpusa izveide ir apjomīgāks uzdevums šajā projektā,
* nav gatavu korpusu valsts pārvaldes jomā, tāpēc tie projektā būs jāizveido, bet korpusu izveidei nepieciešamie teksti eksistē.

### Valodas Korpusu izveide

Lai izveidotu angļu-latviešu MT sistēmu, ir nepieciešams gan angļu-latviešu paralēlais korpuss, gan latviešu monolingvālais korpuss. SMT platformas izmanto paralēlo korpusu, lai no tā izgūtu vārdu un frāžu tulkojumus un to varbūtības, un monolingvālo korpusu, lai no tā izgūtu informāciju par izejas valodas tekstu, piemēram, vārdu un frāžu secību, vārdu savstarpējo saskaņojumu u.tml. Lai izveidotu sākuma līmeņa MT sistēmu šaurā jomā, ir nepieciešami vismaz 1,5 miljoni teikumi (apmēram 30 miljonu vārdu) dotās jomas paralēlais korpuss un vismaz 15 miljonu teikumu monolingvālais korpuss. Augstas kvalitātes MT sistēmas izveidei nepieciešams vairāku desmitu miljonu teikumu liels paralēlais korpuss un vairāku simtu miljonu teikumu liels monolingvālais korpuss. MT izveidei nepieciešamā korpusa apjoms ir atkarīgs arī no valodu specifikas, ja MT sistēmas ieejas un izejas valodas ir līdzīgas un tām ir samērā vienkārša morfoloģija, tad nepieciešams mazāks korpuss, ja valodas ir stipri atšķirīgas, tām ir sarežģīta morfoloģija un būtiski atšķirīga vārdu kārtība teikumā, tad MT izveidei nepieciešams būtiski lielāks korpuss.

MT izveidei un pielāgošanai konkrētajai jomai tiek izmantoti gan konkrētās jomas korpusi, gan vispārīga lietojuma korpusi. Statistiskie modeļi tiek izveidoti gan no vieniem, gan otriem datiem, bet tulkojot šie modeļi tiek lietoti ar dažādiem svariem. Tas ļauj MT sistēmai gan efektīvi pielāgoties konkrētam lietojumam, gan izmantot statistiski ticamus modeļus retāk lietotu vārdu tulkošanai un vispārēja lietojuma tekstu tulkošanai. Svarīgi arī korpusu anotēt ar metadatiem, kas ļautu to vēlāk sašķirot vispārīga lietojuma korpusā un jomai (e-pakalpojumiem) specifiskajā korpusā. Īpaša uzmanība pievēršama valsts pārvaldes tematikai atbilstošu paralēlo dokumentu savākšanai, lai specializētu MT sistēmas tieši šai jomai, t.i., lai sistēmas spētu piemēroties šīs jomas specifiskajai terminoloģijai, valodas konstrukcijām un stilistiskajām īpatnībām.

Projektā jānodrošina 3 tulkošanas virzieni – angļu-latviešu, latviešu-angļu un latviešu-krievu, tāpēc jāvāc gan latviešu-angļu, gan latviešu-krievu paralēlais korpuss, gan monolingvālais korpuss visās 3 valodās.

Ņemot vērā iepriekš minēto, 4.tabula „Projektā nepieciešamie korpusi un to apjoma novērtējums parāda, kādi korpusi ir jāizveido (vai jāapkopo jau esoši) un kādiem ir jābūt šo korpusu apjomiem. Tabulā dots gan minimālais korpusa apjoms, kas ļautu izveidot tikai sākuma līmeņa SMT sistēmu, un bāzes korpusa apjoms, kas ļautu izveidot samērā labas kvalitātes SMT sistēmu. Veicot apjoma novērtējumu, ņemta vērā gan valodu pāra, gan jomas specifika, gan tekstu pieejamība.

4.tabula „Projektā nepieciešamie korpusi un to apjoma novērtējums”

| **Korpuss** | **Joma** | **Korpusa tips** | **Sākuma apjoms**  **(milj. teikumu)** | **Bāzes apjoms (milj. teikumu)** |
| --- | --- | --- | --- | --- |
| angļu-latviešu | vispārēja | paralēlais | 1,5 | 5 |
| krievu-latviešu | vispārēja | paralēlais | 1,5 | 5 |
| angļu-latviešu | valsts pārvalde | paralēlais | 0,3 | 2 |
| krievu-latviešu | valsts pārvalde | paralēlais | 0,3 | 2 |
| Angļu | vispārēja | monolingvālais | 15 | 50 |
| Latviešu | vispārēja | monolingvālais | 25 | 75 |
| Krievu | vispārēja | monolingvālais | 25 | 75 |
| Angļu | valsts pārvalde | monolingvālais | 5 | 15 |
| Latviešu | valsts pārvalde | monolingvālais | 7 | 20 |
| Krievu | valsts pārvalde | monolingvālais | 7 | 20 |

Lai tekstu korpuss būtu izmantojams SMT sistēmu izveidei, tam ir jābūt sagatavotam SMT sistēmu trenēšanai nepieciešamā formātā. Monolingvālajiem korpusiem ir jābūt vienkārša teksta formātā (*plain text*), UTF-8 kodējumā, katram teikumam jābūt jaunā rindā un teikumiem korpusā ir jābūt unikāliem. Paralēlajiem korpusiem, atšķirībā no monolingvālajiem, ir jābūt arī sastatītiem teikumu līmenī, tāpēc paralēlajiem korpusiem ir jābūt kā 2 failiem vienkārša teksta formātā (*plain text*), *UTF-8* kodējumā, katram teikumam jābūt jaunā rindā. Vienā failā ir jābūt vienas valodas teikumiem, otrā – otras valodas teikumiem. Teikumu sastatījums tiek nodrošināts ar secību failos, t.i., abu failu katrā rindiņā ir teikumi, kas ir viens otra tulkojums. Paralēlajā korpusā teikumu pāriem ir jābūt unikāliem. Gan paralēlie, gan monolingvālie korpusi tiek veidoti no tekstiem, kas sākotnēji nav tādā formā kā beigās ir nepieciešams. Izejas teksti var būt gan dokumenti *Microsoft Word, PDF, HTML* vai citos formātos, gan drukāti teksti, piemēram, grāmatas, avīzes u.tml. Ja korpusa izveidē tiek izmantoti drukāti teksti, tad tie ir jāskenē, jāsaglabā kā attēli un jāpārvērš elektroniskos dokumentos, veicot *OCR* procesu. *OCR* procesā ir jānodrošina vismaz 98% atpazīšanas precizitāte. Tā kā izveidojamo korpusu apjoms ir ļoti liels, tad visām tekstu apstrādes darbībām (skenēšana, *OCR,* dažādu dokumentu formātu pārveide par vienkāršu tekstu, teksta dalīšana teikumos, paralēlo tekstu sastatīšana, kvalitātes novērtēšana u.tml.) ir jābūt automatizētām, jo manuāla šādu darbu veikšana ir lēns un ļoti dārgs process. Tā kā automatizētos liela apjoma tekstu apstrādes procesos vienmēr var gadīties dažādas grūti pamanāmas kļūdas un nepilnības, korpusa izveides laikā ir jāsaglabā visi dokumenti (ieskaitot starprezultātus) un programmas, kas tika izmantotas, lai vēlāk būtu iespējams labot kļūdas procesos un atkārtoti izpildīt automātiskos procesus. Tādēļ visi korpusa izveidē izmantotie dokumenti un programmas ir uzskatāmi par korpusa sastāvdaļu.

MT kvalitāte ir galvenokārt atkarīga tieši no tekstu korpusa apjoma un kvalitātes. Eksistē ļoti daudz paralēla teksta, kas būtu noderīgs MT izveidei, taču nodrošināt tā pieejamību var būt sarežģīti. Jāapzinās, ka tulkotie teksti ir izkaisīti, atrodas dažādās vietās, ir dažādos formātos, nav sastatīti, nav mehānismu, kā nodrošināt, ka dažādās institūcijās tulkoti teksti tiek apkopoti u.c. Tāpēc daudzvalodu korpusa vākšana projektā ir prioritārs uzdevums, kura progress nepārtraukti jāizvērtē. Iespējams, ka uzdevuma efektīvai veikšanai nepieciešamas izmaiņas normatīvajos aktos, kas risinātu tekstu pieejamības un lietošanas tiesību jautājumus.

MT sistēmas kvalitāte un efektīvs lietojums ir atkarīgi no vides izmaiņām. Sistēma, kas ir izveidota izmantojot šodien pieejamos tekstu korpusus, vairs nebūs tikpat efektīva nākotnē, jo, laikam ejot mainās valoda, mainās joma, parādās jauni e-pakalpojumi, notikumi, aktualitātes un prioritātes. Ir risks, ka sistēma var novecot. Jāparedz sistēmas uzturēšana ilgtermiņā. Tekstu korpusa atjaunināšanai ir jāturpinās arī pēc sistēmas ieviešanas, un MT sistēma ir regulāri jāpārtrenē izmantojot visus jaunākos datus.

Realizējot projektu tiek plānots atbildību par valodu korpusa izveidošanu un adaptēšanu pilnībā nodot sistēmas izstrādātājam, tādejādi samazinot ar to saistītos riskus un nodrošinot efektīvu līdzekļu izmantošanu.

### MT sistēma

Pašlaik ir pieejami vairāki publiski MT servisi, piemēram *Google Translator, Microsoft Translator*, taču tiešā veidā tie nav izmantojami kā MT servisi e-pārvaldes jomā, jo nav pielāgoti konkrētajai jomai, kā rezultātā tulkojuma kvalitāte nav augsta, un tie nenodrošina vajadzīgās pakalpojuma drošības prasības. Tajā pašā laikā šie servisi var kalpot kā piemēri, kas parāda, kā mašīntulkošana var tikt integrēta dažādos tiešsaistes pakalpojumos.

Ar e-Pārvaldi saistītie pakalpojumi tipiski ir pieejami, izmantojot interneta pārlūkprogrammu, un tos plaši lieto lietotāji, kas savos datoros nav instalējuši nekādu speciālu programmatūru, lai piekļūtu šiem pakalpojumiem. Tāpēc arī MT pakalpojumam ir jābūt pieejamam izmantojot tikai interneta pārlūkprogrammu sniegtās iespējas, un MT servisa nodrošināšanai neder lietotāja datorā instalējama programmatūra, tai skaitā interneta pārlūkprogrammu paplašinājumi jeb pievienojumprogrammas. Ja neapskatām lietotāja datorā instalējamas programmas, tad eksistē 2 klasiski veidi kā MT servisus ieintegrēt interneta vietnēs – 1) izmantojot tulkošanas logrīku (*widget*) 2) izmantojot tulkošanas API.

Izmantojot tulkošanas logrīku, MT integrēšana interneta lapā ir ļoti vienkārša un neprasa būtiskas izmaiņas jau esošajās lapās. Tulkošanas logrīka integrēšana jau esošās ar e-Pārvaldi saistītās lapās neprasa izmaiņas šo lapu funkcionalitātē, šādas integrēšanas veikšanā nav jāiesaista programmatūras izstrādātāji un to var veikt lapas satura redaktori. Piemēram, Latvijas Pašvaldību savienības mājas lapā[[68]](#footnote-69)[[69]](#footnote-70) var redzēt, kā tajā, izmantojot *Microsoft Translator* tulkošanas logrīku[[70]](#footnote-71), ir integrēta MT.

Cits veids kā ieintegrēt MT servisu ir izmantot mašīntulkošanas API. Šis veids ir tehnoloģiski sarežģītāks, tas prasa programmatūras izstrādātāju iesaistīšanos. Taču tas ļauj veidot funkcionāli spēcīgākus risinājumus. Piemēram, šādā veidā integrējot lapas veidotājiem ir pilna kontrole pār to kuras daļas lapā tiks mašīntulkotas, kuras netiks, kad un kādā virzienā teksts tiks tulkots, kā arī ir iespēja izmantot MT pakalpojumu lejuplādējamu dokumentu formēšanā, vai e-pasta vēstuļu sagatavošanā u.tml.

SMT tulkošanas kvalitāte ir atkarīga gan no pielāgojuma valodu pārim, no tekstiem, kas izmantoti SMT sistēmas izveidošanai/trenēšanai, t.i., no pielāgojuma konkrētajai jomai. Vispārēja lietojuma MT sistēmas (tādas kā *Google* un *Microsoft* publiski pieejamās), kas nav pielāgotas ar e-Pārvaldi saistītām jomām, spēj nodrošināt tulkojuma kvalitāti, kas ir pietiekama, lai valodas nezinātājs varētu aptuveni saprast tulkojamā teksta būtību, taču šo sistēmu tulkojuma kvalitāte nav pietiekama, lai tās izmantotu e‑pakalpojumu sniegšanai. Ja MT sistēma ir pielāgota gan valodu pārim, gan konkrētajai jomai, SMT sistēma spēj nodrošināt kvalitātes līmeni, kas ļauj to lietot reālu pakalpojumu sniegšanai, tiesa, ar atsauci, ka teksts ir mašīntulkots un tajā var būt kļūdas. Piemērs pielāgotai MT sistēmai, kas tiek lietota reāla pakalpojuma sniegšanai ir *Microsoft* tehniskās palīdzības zināšanu bāze[[71]](#footnote-72) vācu, japāņu u.c. valodās, kas ir pilnībā mašīntulkota.

### Infrastruktūra

MT servisa izveidei un vēlāk arī darbināšanai ir nepieciešama piemērota infrastruktūra (gan aparatūra, gan programmatūra). Dažādos projekta un tā rezultātu ekspluatācijas posmos ir paredzētas dažādas prasības attiecībā uz infrastruktūru.

Raugoties no infrastruktūras viedokļa var izdalīt 3 projekta procesus, kuriem ir savas atšķirīgas un lielā mērā neatkarīgas prasības attiecībā uz infrastruktūru:

1. valodas korpusu izveide (vākšana, apkopošana un apstrāde, t.i., pārveide SMT trenēšanai nepieciešamajā formātā),
2. SMT sistēmu trenēšana,
3. MT servisa darbināšana.

Projektā ir paredzēti šādi procesi attiecībā uz nepieciešamo infrastruktūru:

1. Tiks veikta detalizēta pieejamās infrastruktūras analīze (VRAA pieejamais datu centrs un tajā izvietotās tehniskās ierīces), lai izvērtētu kādi resursi ir pieejami, kādas iespējas ir tos izmantot ar projektu saistītiem uzdevumiem, un kādi infrastruktūras elementi ir jāveido no jauna.
2. Tiks novērtēta plānotā MT pakalpojuma noslodze (balstoties uz esošajiem [www.latvija.lv](http://www.latvija.lv) un ārpus tās publiski pieejamajiem e-pakalpojumiem un iestāžu mājas lapu skaitu un tajās pieejamo datu apjoma aprēķinu) un tās izmaiņas laikā, lai plānotu optimālo infrastruktūru, kas nodrošinātu optimālu MT pakalpojuma veiktspēju.
3. Tiks izveidots risinājums, kas ļaus dinamiski pārdalīt skaitļošanas resursus starp korpusu izveidei nepieciešamajiem datu apstrādes uzdevumiem, SMT trenēšanas uzdevumiem un SMT pakalpojuma darbināšanu, atkarībā no pieejamajiem resursiem, uzdevumu prioritātes un noslodzes katrā no uzdevumiem.
4. Tiks izveidota nepieciešamā infrastruktūra (aparatūras uzstādīšana, programmatūras instalēšana un konfigurēšana u.tml.).

### Integrēšana citos e-pakalpojumos

Paredzēts izveidot MT servisu, kas būs viegli integrējams e-Pārvaldes sniegtajos e-pakalpojumos.

Paredzēti 2 veidi kā MT pakalpojumus ieintegrēt interneta vietnēs – 1) izmantojot tulkošanas logrīku (*widget*) 2) izmantojot tulkošanas API.

MT servisa veiksmīga darbība ir cieši saistīta ar citiem e-pakalpojumiem. MT serviss nodrošinās tikai tulkošanas pakalpojumu. Lai e-pakalpojumi būtu pieejami vairākās valodās, tajos jaunais MT serviss ir jāintegrē. Risks ir, ka, e-pakalpojumus sniedzošās institūcijas var nebūt ieinteresētas veikt MT integrāciju. Lai mazinātu šo risku MT serviss jāveido maksimāli viegli integrējams, jāizstrādā detalizētas instrukcijas un metodiskas rekomendācijas e-Pārvaldes pakalpojumu izstrādātājiem un uzturētājiem, parādot MT lietošanas ieguvumus, kā arī ir jāizvērtē iespēja veikt nepieciešamās izmaiņas attiecīgajos normatīvajos aktos un vadlīnijās, kas nodrošinātu MT plašu izmantošanu un noteiktu to kā obligātu komponenti atsevišķos valsts pārvaldes risinājumos (mājaslapas, e-pakalpojumi utt.).

### Kvalitātes mērījumi

MT sistēmu tulkošanas kvalitāte ir atkarīga no vairākiem faktoriem – izmantotās tehnoloģijas, pielāgojuma valodu pārim, izmantotajiem tekstu korpusiem u.c. Lai arī MT ir kļuvusi par pasaulē plaši lietotu tehnoloģiju, mūsdienu MT sistēmas nespēj nodrošināt ļoti augtu kvalitāti, salīdzinot ar cilvēka tulkojumu. Regulāri jāveic iegūtās sistēmas kvalitātes novērtējums. Šeit izmantojamas automatizētās kvalitātes novērtēšanas metodes, salīdzinot ar cilvēka veiktu tulkojumu (piem., *BLUE* koeficients) un lietotāju grupas veikts cilvēka novērtējums.

SMT sistēmu trenēšanas process parasti ir iteratīvs, t.i., sistēma tiek trenēta vairākkārt izmantojot dažādus statistiskos modeļus un dažādus sistēmas parametrus. Uztrenētās sistēmas tiek vērtētas gan izmantojot automātiskas MT kvalitātes mērīšanas metodes, piemēram, *BLEU* metriku[[72]](#footnote-73), gan iesaistot sistēmas kvalitātes vērtēšanā cilvēkus, piemēram, izmantojot sistēmu salīdzināšanas metodi[[73]](#footnote-74), gan analizējot uztrenētās sistēmas lietojamību konkrētos lietojumos. Šī procesa rezultātā tiks atrasti optimālie SMT sistēmas modeļi un parametri.

MT sistēmas kvalitāte un efektīvs lietojums ir atkarīgi no vides izmaiņām. Sistēma, kas ir izveidota izmantojot šodien pieejamos tekstu korpusus, vairs nebūs tikpat efektīva nākotnē, jo, laikam ejot mainās valoda, mainās joma, parādās jauni e-pakalpojumi, notikumi, aktualitātes un prioritātes. Ir risks, ka sistēma var novecot. Jāparedz sistēmas uzturēšana ilgtermiņā. Tekstu korpusa atjaunināšanai ir jāturpinās arī pēc sistēmas ieviešanas, un MT sistēma ir regulāri jāpārtrenē izmantojot visus jaunākos datus.

## Plānotajos procesos iesaistītās personas

Plānotais personāls, kas tiks iesaistīts MT izveides procesos ir:

| **Nr.** | **Process** | **Iesaistītās personas** |
| --- | --- | --- |
|  | Jāizveido bāzes MT serviss, izmantojot pieejamās SMT platformas un jau pieejamos paralēlo un monolingvālo tekstu korpusus; | Personāls: IT darbinieki, MT servisa izstrādātāja cilvēkresursi  Iestādes: VA „Kultūras informācijas sistēmas” kā atbildīgais par MT e-pakalpojumu; VISS uzturētājs (dokumenta veidošanas brīdī VRAA) |
|  | Jāpielāgo SMT sistēmas konkrētiem valodu pāriem un e-pakalpojumu specifikai; | Personāls: IT darbinieki, tulkošanas speciālisti, MT servisa izstrādātāja cilvēkresursi  Iestādes: VA „Kultūras informācijas sistēmas” kā atbildīgais par MT e-pakalpojumu; VISS uzturētājs (dokumenta veidošanas brīdī VRAA), projekta sadarbības partneri (piem. LNB) |
|  | Jāizveido MT servisam nepieciešamā infrastruktūra (gan programmatūras, gan aparatūras); | Personāls: IT darbinieki, MT servisa izstrādātāja cilvēkresursi  Iestādes: VA „Kultūras informācijas sistēmas” kā atbildīgais par MT e-pakalpojumu; VISS uzturētājs (dokumenta veidošanas brīdī VRAA) |
|  | MT serviss jāieintegrē e-pakalpojumu vidē. | Personāls: IT darbinieki, MT servisa izstrādātāja cilvēkresursi, to valsts iestāžu darbinieki, kuru infrastruktūrā tiks integrēts MT sīkrīks vai izmantots API e-pakalpojumu tulkošanai.  Iestādes: VA „Kultūras informācijas sistēmas” kā atbildīgais par MT e-pakalpojumu; VISS uzturētājs (dokumenta veidošanas brīdī VRAA) |

Plānotais personāls, klienti un iestādes, kas tiks iesaistīts MT darbināšanā ir tāds pats kā esošajos procesos, skat. 2.2 Esošajos procesos iesaistītās personas.

## Plānotā sistēma

### MT sistēma

Nodaļā „2.6 Mūsdienīgu mašīntulkošanas risinājumu analīze” jau tika minēts, ka pēdējā laikā pasaulē pārsvarā attīsta SMT. SMT izveidošana un pielāgošana specifiskai jomai jeb lietojumam ir daudz lētāka par RBMT. Lai izveidotu vienkāršu SMT sistēmu ir nepieciešamas platformas sistēmu trenēšanai un darbināšanai un liels tekstu korpuss sistēmu trenēšanai. Ir izveidotas labas atvērtā koda platformas SMT trenēšanai un darbināšanai. Tās ir izmantojamas arī latviešu valodai, trenējot uz ļoti liela apjoma latviešu, angļu un krievu paralēlā teksta korpusiem. Tulkošanas kvalitātes uzlabošanai šīm platformām jāizstrādā valodu specifiski moduļi un jāveic citi uzlabojumi, lai pielāgotu konkrētai jomai.

Populārākā un praksē visplašāk lietotā platforma SMT sistēmu trenēšanai un darbināšanai ir *Moses*[[74]](#footnote-75). *Moses* MT rīkkopa tiek izplatīta ar *LGPL* licenci. Tajā ir visi nepieciešamie komponenti, lai veiktu datu priekšapstrādi un trenētu tulkošanas un valodu modeļus. *Moses* MT rīkkopai ir aktīva pētnieku kopiena ar vairāk kā 1000 lietotājiem. Tā ir plaši lietota gan akadēmiskiem pētījumiem, gan komerciālos lietojumos. *Moses* sistēmas lietotāji netiek reģistrēti un nav datu par tās izmantošanu komerciālos nolūkos, tomēr ir zināms ka to lieto šādos nolūkos lieto tādas kompānijas kā *Systran, Asia Online, Autodesk, Matrixware, Translated.net* un Tilde[[75]](#footnote-76).

Tādējādi var secināt, ka *Moses* rīkkopa konceptuāli ir piemērota, lai to izmantotu par pamatu MT servisa izveidei e-Pārvaldes vajadzībām. *Moses* rīkkopa nodrošina ļoti daudzas SMT funkcijas, taču tā nav gatavs risinājums, ko atliek ieviest. Papildus *Moses* iespējām vēl ir nepieciešams izstrādāt:

* programmatūras infrastruktūru, kas ļaus ērti un efektīvi trenēt SMT sistēmas
* MT servisa integrācijai nepieciešamo programmatūru – API, tīmekļa lapu tulkošanas logrīku (*widget*) u.c.
* Pielāgojums tulkošanas virzienam, t.i., valodu pārim, papildinot *Moses* standarta funkcionalitāti ar zināšanām par valodu pāri – morfoloģija, teikuma sintakse u.tml.
* Pielāgojumu jomai, t.i., ePārvaldes un e-pakalpojumu specifikai. MT sistēma ir jāpaplašina, lai tā spētu izmantot jomai specifisku terminoloģiju. Nereti pielāgošana jomai nozīmē arī specifisku teksta iezīmju un marķējuma apstrādi tulkošanas procesā.
* Iespēju iesaistīt MT servisa lietotājus vai administratorus pakalpojuma pielāgošanā un kvalitātes uzlabošanā.

Lietotāji MT servisu galvenokārt neizmantos tiešā veidā, bet izmantos to strādājot ar e‑pakalpojumiem. MT serviss būs integrēts e‑pakalpojumos. Vienkāršākajā integrācijas variantā lietotājs e‑pakalpojuma lapā varēs izvēlēties valodu kurā skatīties lapas saturu. Tipiski e-pakalpojumu lapas ir pieejamas latviešu valodā, ja lietotājs izvēlēsies krievu, vai angļu valodu, tad lapas saturs tiks automātiski tulkots šajās valodās izmantojot MT servisu. MT servisu varēs ieintegrēt arī e-pakalpojumos, kas dažādus dokumentus, izsūta e‑pasta sūtījumus u.tml.

Papildus tam, ka MT serviss būs integrēts e‑pakalpojumos, būs pieejama arī atsevišķa tulkošanas lapa, kurā MT servisa lietotājs vienā logā varēs ievadīt tulkojamo tekstu un otrā logā iegūt tā tulkojumu (līdzīgi kā *Google Translator*[[76]](#footnote-77) vai *Bing Translator*[[77]](#footnote-78) lapās).

MT servisa izstrādē paredzēts izmantot arī gatavas rīkkopas un funkciju bibliotēkas, kas nodrošina bāzes funkcionalitāti, kas projekta gaitā tiks papildināta un uzlabota. Paredzēts SMT trenēšanai un darbināšanai izmantot *Moses* SMT rīkkopu. Izstrādājamo SMT sistēmu pielāgošanai darbam ar latviešu, krievu un angļu valodām paredzēts izmantot gatavus dabīgās valodas apstrādes komponentus, tai skaitā, morfoloģiskos analizatorus un marķētājus, sintaktiskos analizatorus, vārdnīcas.

Paredzēts, ka MT serviss būs pielāgojams katram konkrētam lietojumam. Vispārīgi MT serviss būs pielāgots ePārvaldes vajadzībām, pielāgojot to valodu pāriem un ePārvaldes jomai, taču būs iespēja veikt papildus pielāgojumus, integrējot MT servisu konkrētās e-pakalpojumu interneta lapās. Ja konkrētās lapas uzturētājs pamanīs, ka kādu teikumu, frāzi vai terminu, ko MT serviss tulko slikti, tad būs iespēja norādīt labāku tulkojumu un šis norādījums tiks ņemts vērā tulkojot tekstus.



1.ilustrācija „Risinājuma arhitektūra”

1.ilustrācija parāda, kā paredzēts veidot risinājuma kopējo arhitektūru. MT serviss lietotājiem būs pieejams vairākos veidos – tas būs integrēts e-pakalpojumus izmantojot MT servisa logrīku vai MT servisa API, tas būs pieejams izmantojot MT servisa interneta lapu. MT servisa API nodrošinās unificētu pieeju MT funkcionalitātei, ko nodrošinās ePārvaldes vajadzībām pielāgota SMT sistēma. SMT sistēma izmantos SMT modeļus, kas tiks izveidoti (satrenēti) no paralēlajiem un monolingvālajiem korpusiem.

### Infrastruktūra

MT servisa izveidei un vēlāk arī darbināšanai ir nepieciešama piemērota infrastruktūra (gan aparatūra, gan programmatūra). Dažādos projekta un tā rezultātu ekspluatācijas posmos ir paredzētas dažādas prasības attiecībā uz infrastruktūru.

Projekta sākumā paredzēts veikt detalizētu analīzi par esošo infrastruktūru, lai saprastu tās noslodzi un iespējām to izmantot projekta gaitā un vēlāk rezultātu ekspluatācijā.

Raugoties no infrastruktūras viedokļa var izdalīt 3 projekta procesus, kuriem ir savas atšķirīgas un lielā mērā neatkarīgas prasības attiecībā uz infrastruktūru:

* valodas korpusu izveide (vākšana, apkopošana un apstrāde, t.i., pārveide SMT trenēšanai nepieciešamajā formātā),
* SMT sistēmu trenēšana,
* MT pakalpojuma darbināšana.

Valodas korpusa izveides procesā svarīgas ir 2 galvenās prasības attiecībā uz infrastruktūru – 1) nepieciešama liela apjoma datu glabātuve, kurā tiks glabāti savāktie korpusi un arī visi korpusu apstrādes starprezultāti 2) nepieciešamas lielas skaitļošanas jaudas, korpusu apstrādes uzdevumu veikšanai (piemēram, *OCR*, failu formātu pārveidei, tekstu sastatīšanai u.tml.). Datu glabātuvei ir jābūt pietiekami lielai, lai tā spētu glabāt gan visus izejas failus, gan starprezultātus (tekstu skenēšanas rezultātā iegūtie attēli, *OCR* rezultātā iegūtie dokumenti u.tml.), gan korpusu gala variantu. Svarīgi nodrošināt datu drošību datu glabātuvē, t.i., piekļuves tiesību kontroli, regulāru rezerves kopiju veidošanu u.tml. Plānots, ka valodas korpusu izveides procesā būs nepieciešama datu glabātuve ar vismaz 2 TB ietilpību. Atšķirībā no datu glabātuves, kas procesa laikā būs nepieciešama pastāvīgi, datu apstrādei nepieciešamās skaitļošanas jaudas būs nepieciešamas neregulāri, t.i., datu apstrādes uzdevumi būs jāizpilda neregulāri, bet tad, kad tie būs jāizpilda, tiem būs nepieciešama liela skaitļošanas jauda. Tādējādi ir svarīgi šiem uzdevumiem nodrošināt iespēju pēc pieprasījuma piekļūt lielām skaitļošanas jaudām. Optimāls risinājums ir nodrošināt piekļuvi režģiskās skaitļošanas (*grid computing*) resursiem. Tieši režģiskās skaitļošanas izmantošana šiem uzdevumiem nav obligāta, bet tā kā tā ir svarīga SMT trenēšanas un darbināšanas uzdevumiem, tad, resursu efektīvas izmantošanas nolūkos, to būtu vēlams izmantot arī korpusu izveides darbiem.

SMT sistēmu trenēšanas procesā svarīgas ir 3 galvenās prasības attiecībā uz infrastruktūru – 1) nepieciešama datu glabātuve, kurā tiks glabāti gan satrenētie SMT statistiskie modeļi un sistēmas, gan trenēšanas starprezultāti 2) trenēšanas procesā nepieciešamas lielas skaitļošanas jaudas 3) SMT modeļu trenēšanai ir augstas prasības attiecībā uz operatīvo atmiņu. Tieši tāpat kā korpusu izveides darbiem, arī SMT trenēšanas darbiem skaitļošanas jaudas nebūs nepieciešamas pastāvīgi, jo projekta gaitā SMT sistēmas tiks trenētas epizodiski. Tomēr SMT sistēmas tiks trenētas daudzkārt, jo process būs iteratīvs. SMT sistēmas tiks trenētas izmantojot dažādus statistiskos modeļus un SMT sistēmu parametrus. Visas satrenētās sistēmas tiks izvērtētas, lai atrastu optimālos modeļus un parametrus. Tipiski vienas SMT sistēmas satrenēšanai, atkarībā no modeļu sarežģītības un izmantoto valodas korpusu apjoma, ir nepieciešama 10-30 GB vieta datu glabāšanai un trenēšanas process ilgst 1-3 dienas darbinot to uz viena datora, un trenēšanā izmantotajam datoram ir vajadzīga vismaz 8-16 GB operatīvā atmiņa. Izmantot trenēšanai 1 datoru ir ļoti neoptimāli, jo trenēšana var ilgt vairākas dienas un gaidīšana uz rezultātiem var rada lielas dīkstāves, tāpēc ir ieteicams trenēšanā izmantot vairākus datorus, kas savā starpā sadala trenēšanas uzdevumus un trenēšanu veic paralēli. Izmantojot paralēlus procesus SMT trenēšanā, rezultātu var iegūt ātrāk un ir iespējams pat vienlaicīgi trenēt vairākas SMT sistēmu konfigurācijas ar atšķirīgiem parametriem. Tādējādi ir svarīgi šiem uzdevumiem nodrošināt iespēju pēc pieprasījuma piekļūt lielām skaitļošanas jaudām. Optimāls risinājums ir nodrošināt piekļuvi režģiskās skaitļošanas (*grid computing*) resursiem. Tas ļautu SMT trenēšanas procesiem pēc vajadzības piekļūt skaitļošanas resursiem, kas varētu tikt izmantoti citiem uzdevumiem laikā kad SMT trenēšana nenotiek.

MT servisa darbināšanas procesā svarīgas ir 3 galvenās prasības attiecībā uz infrastruktūru – 1) nepieciešama ātra datu glabātuve, kurā tiks glabāti gan satrenētie SMT statistiskie modeļi un sistēmas 2) nepieciešamas lielas skaitļošanas jaudas 3) SMT darbināšanai ir augstas prasības attiecībā uz operatīvo atmiņu. Infrastruktūra tiks veidota modulāra, lai mainoties MT servisa izmantošanas intensitātei, būtu iespējams viegli palielināt vai samazināt MT servisa veiktspēju. Tipiski vienam serverim, kas darbina SMT sistēmu, atkarībā no SMT sistēmas sarežģītības, ir nepieciešama apmēram 20-30 GB liela ātra datu glabātuve un 8-32 GB operatīvā atmiņa, un šāds serveris spēj tulkot aptuveni 1 teikumu sekundē. Lai MT servisu integrētu ePārvaldes sistēmās ir jānodrošina pietiekami augsta pakalpojuma veiktspēja, un veiktspēju var nodrošināt izvēloties optimālu serveru skaitu SMT sistēmu darbināšanai. Prasības attiecībā pret SMT sistēmu veiktspēju mainās laika gaitā. Servisam kļūstot populārākam tā izmantošanas intensitāte pieaug. Pakalpojuma intensitāte mainās arī dienas laikā, tāpēc ir svarīgi nodrošināt optimālu infrastruktūru, lai SMT sistēmas darbinošie serveri nestāvētu dīkā nakts laikā un pakalpojums nebūtu paralizēts maksimumstundās. Optimāli ir veidot tādu SMT sistēmu darbināšanas infrastruktūru, kas nodrošina pieņemamu veiktspēju vidējas noslodzes apstākļus, bet kas var pēc pieprasījuma piekļūt papildus skaitļošanas resursiem kad tas ir nepieciešams, lai nodrošinātu pieņemamu veiktspēju palielinātas slodzes apstākļos. Tādējādi ir svarīgi šiem uzdevumiem nodrošināt iespēju pēc pieprasījuma piekļūt lielām skaitļošanas jaudām. Optimāls risinājums ir nodrošināt piekļuvi režģiskās skaitļošanas (grid computing) resursiem. Tas ļautu SMT darbināšanas procesiem pīķa stundās pēc vajadzības piekļūt skaitļošanas resursiem, kas varētu tikt izmantoti citiem uzdevumiem (piemēram, SMT trenēšanas uzdevumiem, vai korpusu izveidei nepieciešamajiem tekstu apstrādes uzdevumiem) laikā, kad SMT pakalpojuma noslodze nav ļoti intensīva.

Gan SMT sistēmu trenēšanā gan SMT sistēmu darbināšanā plānots izmantot *Moses* SMT rīkkopu, kas veidos risinājuma kodolu, kas tiks papildināts ar jaunām izstrādēm, kas pielāgos sistēmu valodu pāriem lietojumam ePārvaldes vajadzībām. *Moses* SMT rīkkopa ir paredzēta darbināšanai *Linux* operētājsistēmā, un tas arī nosaka to, ka MT servisa izveidei un darbināšanai būs vajadzīgi serveri ar *Linux* operētājsistēmu. Lai efektīvi trenētu un darbinātu SMT sistēmas ir jānodrošina dažādu uzdevumu paralēla izpilde uz vairākiem serveriem un skaitļošanas jaudu dinamiska piešķiršana SMT sistēmu trenēšanai un darbināšanai. Tipiski šāda paralēla uzdevumu izpilde un dinamiska skaitļošanas resursu iedalīšana tiek veikta izmantojot režģiskās skaitļošanas sniegtās iespējas. *Moses* SMT rīkkopa atbalsta režģiskās skaitļošanas sistēmas *Sun Grid Engine[[78]](#footnote-79)* izmantošanu SMT trenēšanas procesos, tas arī nosaka to, ka, ja projektā tiks izmantots režģiskā skaitļošana, tad tā būs balstīta uz *Sun Grid Engine*.

### Atbalsta stratēģija

MT servisa atbalsta darbības un pakalpojuma līmeņi ietver:

* Infrastruktūras (programmatūras, aparatūras un vides) atbalstu
* Valodu korpusa un tulkošanas kvalitātes nodrošināšanu un uzlabošanu

Infrastruktūras atbalsta līmenim MT servisam ir jābūt ekvivalentam VISS darbības pakalpojumu kvalitātes kritērijiem, jo MT pakalpojums kļūs par integrētu VISS infrastruktūras daļu.

Attiecībā uz valodu korpusa un tulkošanas kvalitātes nodrošināšanu ir nepieciešams noteikt šādus pienākumus:

* Valodu korpusa pastāvīga papildināšana un trenēšana (skat. 4.4 Plānotais tiesiskais regulējums)
* Tulkošanas kvalitātes analīze un automatizētās pārbaudes (skat. 4.1.7 Kvalitātes mērījumi)

Ņemot vērā esošo normatīvo regulējumu attiecībā uz ERAF projektu ilgtspēju tieši projekta iesniedzējs – Valsts aģentūra „Kultūras informācijas sistēmas” būs atbildīgs par sasniegto rezultātu uzturēšanu.

Lai nodrošinātu sistēmas attīstību ilgtermiņā ir jāparedz adekvāti līdzekļi sistēmas uzturēšanai un turpmākai attīstībai, kā arī jāiedibina attiecīgas procedūras. Koncepcijas apraksta veidošanas brīdī ir identificēti vairāki iespējamie ilgtermiņa attīstības risinājumi (piemēram, integrācija ar publisko bibliotēku tīklu publiskajiem interneta piekļuves punktiem u.c.), kas var nodrošināt sistēmas attīstību ar salīdzinoši nelieliem izdevumiem, to sekmīga ieviešana būs tieši atkarīga no sistēmas turētāja aktivitātes un inovatīvās pieejas, kā arī dažādu iestāžu sadarbībspēju. Nepieciešamais finansējums būs tieši atkarīgs no izvēlētajām procedūrām un to izmaksām, kā arī tehniskās infrastruktūras uzturēšanas izmaksām, tomēr pašlaik var secināt, ka kopumā uzturēšanai būs nepieciešams viens štata darbinieks uz pilnu slodzi, tiek plānots, ka šāds darbinieks būs nepieciešams sākot ar 2012.gadu.

## Plānotais tiesiskais regulējums

Plānotās izmaiņas normatīvajos aktos ir saistītas ar nepieciešamību iekļaut papildinājumus, kas nodrošinātu tehniskās iespējas MT iekļaušanai jaunos e-pakalpojumos kā arī ieteicams veikt uzlabojumus.

| **Normatīvais akts** | **Izmaiņu būtība un ietekme** | **Rīcība** |
| --- | --- | --- |
| 2007.gada 6.marta MK noteikumi Nr.171 „Kārtība, kādā iestādes ievieto informāciju internetā”[[79]](#footnote-80) | Ieteicams papildināt normatīvos aktus ar prasību izmantot izstrādāto MT logrīku automātiskās tulkošanas nodrošināšanai.  Šādā veidā uzlabosies MT lietojums, būs iespēja uzlabot tulkojumu kvalitāti, samazināsies izmaksas tulkojumu nodrošināšanai un uzlabosies informācijas pieejamība citās valodās. | Virzīt grozījumus pēc MT logrīka izstrādes. |
| Obligāto eksemplāru likums | Ieteicams nodrošināt obligātu elektronisku satura nodošanu Latvijas Nacionālai bibliotēkai vai citam MT satura turētājam arī iespieddarbu tulkojumiem.  Šādā veidā tiks nodrošināta valodu korpusa pieejamību MT tulkojumu kvalitātes uzlabošanai | Virzīt grozījumus pēc MT logrīka izstrādes. |
| Jauns normatīvais akts | MK noteikumu līmenī ir jāparedz projekta ietvaros izveidotās mašīntulkošanas sistēmas un izveidotā elektroniskā pakalpojuma obligātu darbību un lietošanu attiecīgajās valsts pārvaldes informācijas sistēmās, kā arī izmantojot publisko finansējumu tulkoto dokumentu elektronisko versiju pieejamības nodrošināšanu MT valodu korpusam.  Šādā veidā tiks nodrošināta valodu korpusa pieejamība MT tulkojumu kvalitātes uzlabošanai, kā arī izstrādāto risinājumu plaša izmantošana publiskās pārvaldes sistēmās | Izstrādāt un virzīt apstiprināšanai atbilstošu normatīvo aktu pēc sistēmas izstrādes pabeigšanas. |

## Plānotie politikas dokumenti

MT ieviešana neprasa veikt izmaiņas politikas dokumentos.

# Plānotās sistēmas konceptuālā risinājuma analīzes kopsavilkums

## Citu valstu pieredzes analīze

Informācijas un satura pakalpojumu apjomam nepārtraukti pieaugot, tulkotāju piesaiste plaša apjoma tulkošanai kļūst nesamērīgi dārga, tāpēc pasaulē un ES arvien plašāk sāk izmantot MT tehnoloģijas. Eiropas Komisija nodrošina MT servisu valsts pārvaldes iestādēm, diemžēl tas ir pieejams tikai lielajām Eiropas valodām[[80]](#footnote-81). MT servisi publisko pakalpojumu nodrošināšanai tiek izstrādāti un nodrošināti arī vairākām mazākām ES valodām, piemēram, basku-spāņu[[81]](#footnote-82) vai angļu-lietuviešu[[82]](#footnote-83). Pēdējā laikā pie MT pakalpojumu ieviešanas strādā arī Eiropas Patentu Birojs[[83]](#footnote-84), kas vienotā Eiropas patenta ietvaros plāno izmantot MT pakalpojumus, lai tulkotu patentus no angļu valodas uz citām Eiropas valodām[[84]](#footnote-85).

## Analīze par mašīntulkošanas servisu integrāciju e-pārvaldes infrastruktūrā

Ar ePārvaldi saistītie pakalpojumu tipiski ir pieejami izmantojot interneta pārlūkprogrammu, un tos plaši lieto lietotāji, kas savos datoros nav instalējuši nekādu speciālu programmatūru, lai piekļūtu šiem pakalpojumiem. Tāpēc arī MT pakalpojumam ir jābūt pieejamam izmantojot tikai interneta pārlūkprogrammu sniegtās iespējas, un MT pakalpojuma nodrošināšanai neder lietotāja datorā instalējama programmatūra, tai skaitā interneta pārlūkprogrammu paplašinājumi jeb pievienojumprogrammas. Ja neapskatām lietotāja datorā instalējamas programmas, tad eksistē 2 klasiski veidi kā MT servisus ieintegrēt interneta vietnēs – 1) izmantojot tulkošanas logrīku (*widget*) 2) izmantojot tulkošanas API.

Paredzēts, ka MT serviss būs viegli integrējams ePārvaldes sniegtajos e-pakalpojumos.

Izmantojot tulkošanas logrīku, MT integrēšana interneta lapā ir ļoti vienkārša un neprasa būtiskas izmaiņas jau esošajās lapās. Tulkošanas logrīka integrēšana jau esošās ar ePārvaldi saistītās lapās neprasa izmaiņas šo lapu funkcionalitātē, šādas integrēšanas veikšanā nav jāiesaista programmatūras izstrādātāji un to var veikt lapas satura redaktori. Piemēram, Latvijas Pašvaldību savienības mājas lapā[[85]](#footnote-86) var redzēt, kā tajā, izmantojot *Microsoft Translator* tulkošanas logrīku[[86]](#footnote-87), ir integrēts MT pakalpojums.

Cits veids kā ieintegrēt MT servisu ir izmantot mašīntulkošanas API. Šis veids ir tehnoloģiski sarežģītāks, tas prasa programmatūras izstrādātāju iesaistīšanos. Taču tas ļauj veidot funkcionāli spēcīgākus risinājumus. Piemēram, šādā veidā integrējot lapas veidotājiem ir pilna kontrole pār to, kuras daļas lapā tiks mašīntulkotas, kuras netiks, kad un kādā virzienā teksts tiks tulkots, arī iespēja izmantot MT pakalpojumu lejuplādējamu dokumentu formēšanā, vai e-pasta vēstuļu sagatavošanā u.tml.

Paredzēts, ka MT serviss būs pielāgojams katram konkrētam lietojumam. Vispārīgi MT pakalpojums būs pielāgots ePārvaldes vajadzībām, pielāgojot to valodu pāriem un ePārvaldes jomai, t.i., sistēmas izveidē tiks izmantots ePārvaldes jomas korpuss. Taču būs iespēja veikt papildus pielāgojumus, integrējot MT servisu konkrētās e-pakalpojumu interneta lapās. Ja konkrētās lapas uzturētājs pamanīs, ka kādu teikumu, frāzi vai terminu MT serviss tulko slikti, tad būs iespēja norādīt labāku tulkojumu un šis norādījums tiks ņemts vērā tulkojot tekstus.

## Kopsavilkums par plānoto informācijas sistēmu

Plānotais MT risinājums nodrošinās vieglu un ērtu elektroniskās informācijas pieejamību vairākās valodās, bez nepieciešamības lietotājam izmantot speciālu programmatūru. MT risinājums tiks izmantots gan e-Pakalpojumu (oficiāli [www.latvija.lv](http://www.latvija.lv) publicēto un arī iestāžu pašu uz savām tehniskajām platformām nodrošināto) pieejamības uzlabošanai, gan publiskojamās informācijas pieejamības uzlabošanai (nodrošinot iespēju integrēt MT logrīku gan iestāžu mājas lapās, gan citās tīmekļa sistēmās). Plānotais risinājums atvieglos tieši aktuālās informācijas pieejamību dažādām lietotāju grupām, kuras tulkošana neatmaksājas vai to nav iespējams operatīvi veikt.

Ieviešot MT risinājumu kā daļu no e-Pakalpojumu infrastruktūras, tiks nodrošināta iespēja nākotnē vienkāršoti pievienot papildus valodu pārus e-Pakalpojumu un citas informācijas tulkošanai dažādās valodās.

MT serviss, ar valsts pārvaldei pielāgotu valodu korpusu uzlabotai tulkojuma kvalitātei, būs pieejams divos veidos:

* Kā logrīks, ar kura palīdzību varēs tulkot jebkuru tīmekļa lappusi, datu ievades formu vai e-pakalpojuma rezultātu (ar nosacījumu, ka tie ir attēloti *HTML* formātā);
* Kā MT servisa API, kas pieļaus padziļinātu tā integrāciju esošajās informācijas sistēmās (t.sk. publicēto dokumentu tulkošanai u.c. vajadzībām).

Projekta rezultātā izveidotās sistēmas galvenie lietotāji būs iestādes un to klienti – ministrijas un pašvaldības, kā arī to padotības iestādes un kapitālsabiedrības, uz kurām ir attiecināms esošais normatīvais regulējums par informācijas sniegšanu iedzīvotājiem.

Par sistēmas izmantošanu netiek plānots saņemt maksājumus no tās lietotājiem.

Galvenie ieguvumi, realizējot plānoto MT servisu būs:

1. Uzlabojums esošo e-pakalpojumu pieejamībā un plašāka iespēja lietot e-pakalpojumus gan Latvijas iedzīvotājiem, gan citiem interesentiem, nodrošinot to sniegšanu dažādās valodās;
2. Plašāka sabiedrības informētība par valsts pārvaldes procesiem un augstāka sabiedrības līdzdalība dažādos procesos
3. Iespēja sabiedrībai iegūt un lasīt informāciju citās valodās izmantojot MT servisu.

# Galvenie riski/iespējas

Pievienotajā tabulā (5.tabula „Galvenie MT riski/iespējas”) ir apkopota koncepcijas veidošanas brīdī pieejamā informācija par dažādiem pozitīviem (iespējām) un negatīviem riskiem attiecībā uz MT izveidi:

5.tabula „Galvenie MT riski/iespējas”

| **Novērojums** | **Izrietošais risks / iespēja** | **Iespējamā rīcība riska mazināšanai/iespēju izmantošanai** |
| --- | --- | --- |
| Vairāki esošie e-pakalpojumi atgriež rezultātu elektroniski parakstīta elektroniskā dokumenta formātā. Līdz ar to rezultāta dokuments nav maināms nezaudējot tā derīgumu un tā saturs var būt dinamisks (arī vairāki citi dokumenti). | Samazināts MT lietojums e-pakalpojumu vidē tiešās informācijas iegūšanai. | MT rezultātam nav juridiska spēka, kas nozīmē, ka tulkojumam ir informatīvs raksturs. Līdz ar to lietotājam tāpat saglabāsies iespēja veikt tulkojumu izmantojot MT publisko tulkošanas lapu, izgriežot un ielīmējot saturu no pakalpojuma rezultātā iegūtā dokumenta. |
| Vairāki esošie e-pakalpojumi atgriež rezultātu PDF vai RTF datņu formātā jau no orģinālā avota. Līdz ar to var būt traucēta iespēja identificēt e-pakalpojuma saturu strukturētā formā. | Samazināts MT lietojums e-pakalpojumu vidē tiešās informācijas iegūšanai. | MT rezultātam nav juridiska spēka, kas nozīmē, ka tulkojumam ir informatīvs raksturs. Līdz ar to lietotājam tāpat saglabāsies iespēja veikt tulkojumu izmantojot MT publisko tulkošanas lapu, izgriežot un ielīmējot saturu no pakalpojuma rezultātā iegūtā dokumenta. |
| Citus e-pakalpojumus sniedzošās institūcijas var nebūt ieinteresētas veikt MT integrāciju | MT pakalpojums netiks lietots | MT pakalpojums tiks veidots maksimāli viegli integrējams. Tiks izstrādātas detalizētas instrukcijas un metodiskas rekomendācijas ePārvaldes pakalpojumu izstrādātājiem un uzturētājiem (iespējams papildināt IVIS ePakalpojumu standartu) |
| Ir liels skaits ārpus [www.latvija.lv](http://www.latvija.lv) esošo e-pakalpojumu (t.sk. veidlapu un vienkārši publicētas informācijas), kas ir realizēti, izmantojot dažādas tehnoloģijas. | Šim saturam netiks nodrošināts tulkojums. | MT logrīks un MT API būs brīvi pieejams visām iestādēm tā integrācijai savos e-pakalpojumos vai mājas lapās. Logrīka realizācijas mehānisms pieļauj tā integrāciju ar ļoti dažādās tehnoloģijās realizētiem e-pakalpojumiem.  Visām iestādēm tiks izsūtīta informācija par MT servisu pēc tā palaišanas |
| Daudz plašāka informācijas pieejamība citās valodās un citu valstu publicētās informācijas pieejamība latviešu valodā (*iespēja*) | Daudz plašāks pieejamais kvalitatīvi iztulkotas informācijas apjoms gan cittautiešiem par Latviju, gan latviešiem par citām valstīm. | MT servisa publiski pieejamā daļa dos iespēju iztulkot citu valstu dokumentus vai citā valodā runājošajiem tulkot latviski esošos dokumentus, izmantojot speciāli oficiālai saziņai pielāgotu valodu korpusu (ar augstāku tulkojumu kvalitāti) |
| Aktuālās informācijas augsta pieejamība citās valodās (*iespēja*) | Daudz plašāks pieejamais kvalitatīvi iztulkotas informācijas apjoms gan cittautiešiem par Latviju, gan latviešiem par citām valstīm. | Šobrīd vairums valsts iestāžu mājas lapu tulko tikai starptautiskos jaunumus vai statiskās lapas daļas. MT logrīks dod iespēju citās valodās runājošajiem iepazīties ar visaktuālāko informāciju. |

# plānotās sistēmas izveides un uzturēšanas izmaksas un to efektivitātes analīze

## Plānotās sistēmas izveides izmaksas

|  |  |  |
| --- | --- | --- |
| **Pozīcija** | **Apraksts** | **Summa, Ls ar PVN** |
| Infrastruktūras iegādes izmaksas | Tiks iegādāta projekta vajadzībām atbilstoša IT infrastruktūra un nodrošināta tās pilnvērtīga darbība | 50.000,00 |
| Valodu korpusa un MT sistēmas izstrādes izmaksas | Tiks izveidots projekta prasībām atbilstošs valodu korpuss un izstrādāta MT sistēma | 660.000,00 |
| E-pakalpojumu sistēmas uzlabojumi un sistēmas lietojumrisinājumi (sīkrīks u.c.) | Tiks veiktas nepieciešamās izmaiņas esošajos e-pakalpojumos, kā arī izstrādāti atbilstoši risinājumi integrācijai valsts pārvaldes mājas lapās | 50.000,00 |
| Konsultantu un ekspertu izmaksas | Konsultantu un ekspertu piesaistes izmaksas kvalitātes kontroles nodrošināšanai (sistēmas izstrāde gaitas kontrole, e-pakalpojumu uzlabojumu veikšanas kontrole, tehnisko specifikāciju izstrāde u.c.) | 40.000,00 |
| **Kopā:** | | **800.000,00** |

**6.tabula „Plānotās sistēmas izveides izmaksas”**

## Plānotie sistēmas izveides un uzturēšanas finansējuma avoti pa gadiem

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Avots** | **Gads** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** |
| ERAF | LVL (1000) | 290 | 510 |  |  |  |  |  |
| Valsts budžets | LVL (1000) |  | 30 | 80 | 130 | 80 | 80 | 130 |

**7.tabula „Plānotie sistēmas izveides un uzturēšanas izmaksu avoti pa gadiem”**

## Alternatīvu analīze

Attiecībā uz situāciju alternatīvu salīdzinājumu starp stāvokļiem „ar projektu” un „bez projekta”, šajā dokumentā tika analizēta tikai situācija „ar projektu”, jo uz dokumenta sagatavošanas brīdi nav izveidota pastāvoša IS, kas nodrošinātu attiecīgā pielietojuma MT funkcionalitāti, līdz ar to sabiedrība negūst attiecīgo labumu un tas nerada finanšu izmaksas.

Attiecībā uz elektroniskās pārvaldes risinājumu izmantošanas scenārijiem tika izskatīta tikai scenārijs, saskaņā ar kuru elektroniskās pārvaldes komponentes izmantotas netiek, jo pēc būtības šīs Sistēma ir paredzēta, kā viena (papildu) no attiecīgajām komponentēm, kuru nākotnē koplietos citu sistēmu ietvaros.

Pamatojoties uz iepriekš minētajiem apsvērumiem, šajā dokumentā netiek veikta alternatīvu analīze, savukārt, projekta izmaksu efektivitātes analīze tiek veikta tikai vienai alternatīvai – „ar projektu”.

## Projektu izmaksu efektivitātes analīze

Šajā dokumenta sadaļā aprakstītās projekta izmaksu efektivitātes analīze, kuras atspoguļojums sniegts tabulā Nr. 8.

Izmaksu efektivitātes analīzes naudas plūsma tika veikta, pamatojoties uz šādiem pieņēmumiem attiecībā uz darbības izmaksām (skat. atbilstoši kopsavilkuma tabulā norādītajiem numuriem):

1. Papildu infrastruktūras izmaksas, palielinoties pieprasījumam pēc pakalpojuma un nokalpojušās infrastruktūras komponenšu aizvietošana - attiecīgās izmaksas ik trīs gadus būs LVL 50 000 apmērā no sākotnējās kopējās infrastruktūras cenas. Cenas pieaugums netiek paredzēts, ņemot vērā, ka tehniskā nodrošinājuma cenai ir vispārēja tendence samazināties;
2. Programmatūras uzturēšana – attiecīgās izmaksas veidos darbi, kas saistīti ar programmatūras uzturēšanu, kļūdu labošanu funkcionalitātē vai tās atjaunošanu. Paredzēts, ka pēc garantijas perioda beigām izmaksas veidos ārpakalpojumu sniedzēja viena augsti kvalificēta darbinieka slodze, kas izmaksās iestādei LVL 50 000 gadā. Minētās izmaksas tiks finansētas no valsts budžeta;
3. MT iekļaušana jaunos e-pakalpojumos un plašs izveidotās MT funkcionalitātes lietojums – neveidos papildus izmaksas, jo:

1) tulkošanas logrīka ieviešana ir ļoti vienkārša un neprasa būtiskas izmaiņas jau esošajās lapās. Tulkošanas logrīka integrēšana jau esošās ar e-Pārvaldi saistītās lapās neprasa izmaiņas šo lapu funkcionalitātē, šādas integrēšanas veikšanā nav jāiesaista programmatūras izstrādātāji un to var veikt lapas satura redaktori;

2) tulkošanas API ieviešanas attiecīgās darbības veiks attiecīgo mājas lapu īpašnieki.

1. Valodu korpusa pastāvīga papildināšana un trenēšana un tulkošanas kvalitātes analīze un automatizētās pārbaudes – paredzams, ka šīs funkcijas tiks organizētas tādējādi, ka lielāko darbu apjomu veiks Sistēmas lietotāji, kuriem būs pienākums/iespēja sniegt priekšlikumus tulkojumu uzlabošanai, bet viens profesionāls tulks, kura algas izmaksas veidos LVL 25.000 gadā veiks šo lietotāju piedāvāto korekciju analīzi un apstiprināšanu. Lai noteiktu uzturēšanas izmaksu sadārdzinājumu, algai attiecīgi paredzētas ilgtermiņa darba algu palielinājuma prognozes;
2. Attiecībā uz telpu, telekomunikācijas, elektrības izmaksām un administrācijas izmaksām, tiek paredzēts, ka šīs izmaksas tiks ietvertas esošajā IVIS platformas uzturēšanas tehniskajā un cilvēkresursu izmaksu kopumā, līdz ar to neradot papildus izmaksas;
3. Ieņēmumi – ieņēmumu daļa nav paredzēta, šis pakalpojums būs daļa no bezmaksas pieejamajiem IVIS koplietošanas pakalpojumiem.

Izpētes procesā tika identificēti divi galvenie sociālekonomiskie ieguvumi, kuri tika novērtēti, pamatojoties uz tālāk tekstā aprakstīto pieeju.

**Iedzīvotāju neklātienes apkalpošana pretstatā klātienei**

Sistēmas ieviešana nodrošinātu iespēju plašākam sabiedrības lokam izmantot e-pakalpojumus, nodrošinot to kvalitatīvas tulkošanas iespēju.

Sociālekonomiskais ieguvums iestādēm un pašvaldībām

Paredzams, ka e-pakalpojumu plašāka izmantošana samazinās nepieciešamību pēc valsts iestāžu darbiniekiem, kuri šobrīd apkalpo klātienē klientus, taču tai pat laikā palielinās to darbinieku skaitu, kuri veiks e-pakalpojumu pieprasījumu apkalpošanu. Līdz ar to nav paredzams, ka attiecīgās izmaiņas sniegs darbaspēka izmaksu ietaupījumu iestādēm vai pašvaldībām.

Sociālekonomiskais ieguvums iedzīvotājiem

Sistēmas ieviešana nodrošinātu iespēju veikt attālinātu e-pakalpojumu saņemšanu personām, kuras nepārvalda latviešu valodu, kā rezultātā e-pakalpojuma izmantošana ļautu novirzīt laiku produktīvai preču vai pakalpojumu ražošanai.

Attiecīgais ieguvums var tikt izteikts ar šādu formulu:

**,** kur

**A** – papildus iedzīvotāju skaits, kuriem tiks nodrošinātas iespējas izmantot e-pakalpojumus un kas tos izmantos, ko nosaka šādi:

E, kur

D – e-pakalpojumu izmantošanas reizes 2010. gadā [www.latvija.lv](http://www.latvija.lv) mājas lapā[[87]](#footnote-88);

G – Latvijas iedzīvotāju apjoms, kuriem latviešu valoda nav pirmā valoda (atbilstoši Latviešu valodas aģentūras datiem)[[88]](#footnote-89);

H – procentuāli izteikts to iedzīvotāju apjoms, kuru pirmā valoda nav latviešu valoda un līdz ar to nelieto e-pakalpojumus, bet kas vēlas tos izmantot (pieņēmumu skatīt tālāk tekstā);

F – ekonomiski aktīvo iedzīvotāju skaits, kas ietver visus iedzīvotājus, kam Latviešu valoda ir pirmā valoda un pusi to iedzīvotāju, kuriem tā nav pirmā valoda (pieņēmumu skatīt tālāk tekstā);

E – ikgadējais e-pakalpojumu lietotāju pieauguma koeficents (pieņēmumu skatīt tālāk tekstā);

**B** – vidējais IKP apjoms, ko saražo viens iedzīvotājs vienā stundā, ko nosaka šādi:

**C** – laiks, ko ietaupa iedzīvotājs, saņemot e-pakalpojumu attālināti;

Veicot attiecīgos aprēķinus tiek izdarīti šādi pieņēmumi:

* tiek pieņemts, ka tikai 50% no ekonomiski aktīvajiem iedzīvotājiem (ekonomiski aktīvie iedzīvotāji, kas ir sākot ar 18 gadus veci - 83.4%[[89]](#footnote-91)), kuriem latviešu valoda nav pirmā valoda vēlas izmantot e-pakalpojumus, bet valodas ierobežojumu dēļ – nespēj (mainīgās vērtības – H un F);
* tiek pieņemts, ka e-pakalpojumu lietotāju skaits palielināsies lineāri, atbilstoši pagājušā gada rādītājiem (šis ir ļoti piesardzīgs pieņēmums, jo pieejamo e-pakalpojumu skaits tuvākajos gados laiks būtiski pieaugs salīdzinājumā ar jau esošo klāstu) un tas sastādīs 15% ik gadu (mainīgā vērtība – E);
* laiks, ko ietaupa iedzīvotājs, saņemot e-pakalpojumu attālināti tiek pieņemts vismaz 30 minūšu apjomā (mainīgā vērtība – C).

Tulkošanas automatizācija valsts iestādēs

Sistēmas ieviešana nodrošinātu iespēju valsts un pašvaldības iestādēm aktuālās informācijas publicēšanu citās valodās veikt automatizēti, tādējādi ietaupot laiku personām, kas nodrošina šos tulkojumus. Jaunākā tulkošanas atbalsta programmatūra (piem., SDL TRADOS 2009, Kilgray MemoQ u.c.) ietver iespēju tajā integrēt arī automatizēto tulkošanu. Pētījumi rāda (sk. piem., „Proceedings of the 15th conference of the European Association for Machine Translation”), ka šādā veidā tulkotāja darba produktivitāte pieaug vismaz par trešdaļu, kas ļauj būtiski samazināt tulkošanai nepieciešamo laiku un attiecīgi arī izmaksas. Ņemot vērā, ka pašlaik šādi tulkojumi praktiski netiek veidoti un publicēti, tad nav iespējams noteikt pamatotu sociālekonomisko ieguvumu, ko sniegtu šīs informācijas sniegšana un publicēšana.

**8.tabula „Projekta izmaksu efektivitātes analīze”**



Kultūras ministre Ž.Jaunzeme – Grende

Vīza: Valsts sekretāra p.i. U.Lielpēters

28.12.2011. 14:58

11628

A.Magone

Tālr. 67844894; fakss 67843084

[Armands.Magone@kis.gov.lv](mailto:Armands.Magone@kis.gov.lv)

1. European Commission Machine Translation Service: <https://webgate.ec.europa.eu/mt/ecmt/> [↑](#footnote-ref-2)
2. <http://ixa.si.ehu.es/openmt2/> [↑](#footnote-ref-3)
3. <http://vertimas.vdu.lt/> [↑](#footnote-ref-4)
4. K. Papineni, S. Roukos, T. Ward, W. Zhu, BLEU: a method for automatic evaluation of machine translation, in Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL), 2002 [↑](#footnote-ref-5)
5. <http://www.gnu.org/licenses/lgpl.html> [↑](#footnote-ref-6)
6. http://tpi.mk.gov.lv/ui/default.aspx?MenuID=1 [↑](#footnote-ref-7)
7. <http://www.vraa.gov.lv/lv/epakalpojumi/viss/> [↑](#footnote-ref-8)
8. http://www.likumi.lv/doc.php?id=50601 [↑](#footnote-ref-9)
9. http://www.likumi.lv/doc.php?id=154198&from=off [↑](#footnote-ref-10)
10. http://www.likumi.lv/doc.php?id=164501 [↑](#footnote-ref-11)
11. http://www.likumi.lv/doc.php?id=14740 [↑](#footnote-ref-12)
12. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2008:320:0001:01:EN:HTML [↑](#footnote-ref-13)
13. http://lsif.lv/files/pics/Par\_mums/VPSIL-pilna-versija-labais.pdf [↑](#footnote-ref-14)
14. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0566:FIN:LV:PDF [↑](#footnote-ref-15)
15. <http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf> [↑](#footnote-ref-16)
16. Hutchins J. (2006) Machine translation: history of research and use. In Encyclopedia of Languages and Linguistics. 2nd edition, edited by Keith Brown (Oxford: Elsevier 2006), vol.7, pp.375-383. [↑](#footnote-ref-17)
17. <http://www.hutchinsweb.me.uk/Compendium-15.pdf> [↑](#footnote-ref-18)
18. <http://translate.google.com> [↑](#footnote-ref-19)
19. <http://www.microsofttranslator.com> [↑](#footnote-ref-20)
20. <http://translate.tilde.lv> [↑](#footnote-ref-21)
21. <http://www.hutchinsweb.me.uk/Westminster-2009.pdf> [↑](#footnote-ref-22)
22. <http://www.cse.ust.hk/~dekai/library/WU_Dekai/WU_Dekai_MTSummit_20030726.pdf>, <http://www.hutchinsweb.me.uk/CUHK-2006.pdf>, <http://www.mt-archive.info/Translingual-Europe-2009-Uszkoreit-2-ppt.pdf>. [↑](#footnote-ref-23)
23. Brown P., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Mercer R., Roossin P. (1988b) A statistical approach to language translation in Proceedings of the 12th International Conference on Computational Linguistics COLING’88, 22-27 August 1988, (1): 71-76. [↑](#footnote-ref-24)
24. Hutchins J. (2007a) Machine translation in Europe and North America: current state and future prospects. In JAPIO 2007 Yearbook (Tokyo: Japan Patent Information Organization). [↑](#footnote-ref-25)
25. Koehn P., Och F. J., Marcu D. (2003) Statistical phrase based translation. In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL). [↑](#footnote-ref-26)
26. Chiang D. (2007) Hierarchical Phrase-Based Translation. In Computational Linguistics 33(2): 201-228. [↑](#footnote-ref-27)
27. Och F.J. and Ney H. (2004) The alignment template approach to statistical machine translation. In Computational Linguistics, 30:417–449. [↑](#footnote-ref-28)
28. Vogel S., Zhang Y., Huang F., Tribble A., Venugopal A., Zhao B., Waibel A. (2003) The CMU Statistical Machine Translation System. In Proceedings of MT-Summit IX. [↑](#footnote-ref-29)
29. Tillmann C. (2003) A projection extension algorithm for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural language processing, pp. 1-8. [↑](#footnote-ref-30)
30. Koehn P. and Hoang H. (2007) Factored Translation Models. In Proceedings of EMNLP’07. [↑](#footnote-ref-31)
31. Yamada K. and Knight K. (2001) A Syntax-Based Statistical Translation Model. In Proceedings of the Conference of the Association for Computational Linguistics, ACL’02; Yamada K. and Knight K. (2002) A Decoder for Syntax-Based Statistical MT. In Proceedings of the Conference of the Association for Computational Linguistics, ACL’02; May J. and Knight K. (2007) Syntactic Re-Alignment Models for Machine Translation. In Proceedings of EMNLP-CoNLL’07. [↑](#footnote-ref-32)
32. Moses MT rīkkopa (<http://www.statmt.org/moses/>), Giza++ tekstu sastatīšanas rīkkopa (<http://www.fjoch.com/GIZA++.html>), SRILM valodas modelēšanas rīkkopa (<http://www-speech.sri.com/projects/srilm/>) u.c. [↑](#footnote-ref-33)
33. http://www.euromatrixplus.eu [↑](#footnote-ref-34)
34. <http://wt.jrc.it/lt/Acquis/> [↑](#footnote-ref-35)
35. <http://langtech.jrc.it/DGT-TM.html> [↑](#footnote-ref-36)
36. <http://opus.lingfil.uu.se/> [↑](#footnote-ref-37)
37. <http://opus.lingfil.uu.se/EMEA.php> [↑](#footnote-ref-38)
38. <http://www.vvc.gov.lv/> [↑](#footnote-ref-39)
39. <http://www.accurat-project.eu/> [↑](#footnote-ref-40)
40. <http://www.meta-net.eu/events/meta-forum-2010> [↑](#footnote-ref-41)
41. <http://www.eurotermbank.com> [↑](#footnote-ref-42)
42. <http://www.tausdata.org/> [↑](#footnote-ref-43)
43. CLARIN projekts (http://www.clarin.lv) [↑](#footnote-ref-44)
44. <http://www.anc.org/> [↑](#footnote-ref-45)
45. <http://en.wikipedia.org/wiki/Bank_of_English> [↑](#footnote-ref-46)
46. <http://www.natcorp.ox.ac.uk/> [↑](#footnote-ref-47)
47. <http://www.americancorpus.org/> [↑](#footnote-ref-48)
48. <http://www.ucl.ac.uk/english-usage/ice/> [↑](#footnote-ref-49)
49. <http://www.oxforddictionaries.com/page/oec> [↑](#footnote-ref-50)
50. [http://www.statmt.org/wmt09/training-monolingval.tar](http://www.statmt.org/wmt09/training-monolingual.tar) [↑](#footnote-ref-51)
51. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T13> [↑](#footnote-ref-52)
52. <http://www.statmt.org/europarl/> [↑](#footnote-ref-53)
53. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13> [↑](#footnote-ref-54)
54. <http://www.ruscorpora.ru> [↑](#footnote-ref-55)
55. <http://www.ruscorpora.ru/corpora-structure.html> [↑](#footnote-ref-56)
56. <http://www.ruscorpora.ru/search-main.html> [↑](#footnote-ref-57)
57. <http://www.narusco.ru/index.htm> [↑](#footnote-ref-58)
58. <http://www.narusco.ru/project.htm> [↑](#footnote-ref-59)
59. <http://www.slaviska.uu.se/ryska/corpus-shortdescription.html> [↑](#footnote-ref-60)
60. <http://www.philol.msu.ru/~lex/corpus> [↑](#footnote-ref-61)
61. <http://lexicol.philol.msu.ru> [↑](#footnote-ref-62)
62. <http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html> [↑](#footnote-ref-63)
63. <http://www.coli.uni-saarland.de/~thorsten/tnt> [↑](#footnote-ref-64)
64. <http://www.comp.leeds.ac.uk/ssharoff/publications/cl-2003.pdf> [↑](#footnote-ref-65)
65. <http://www.comp.leeds.ac.uk/ssharoff/bokrcorpora/index-en.html> [↑](#footnote-ref-66)
66. <http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf> [↑](#footnote-ref-67)
67. <http://corpus.leeds.ac.uk/internet.html> [↑](#footnote-ref-68)
68. Rakstzīmju optiskā pazīšana (*optical character recognition)* [↑](#footnote-ref-69)
69. <http://www.lps.lv/> [↑](#footnote-ref-70)
70. <http://www.microsofttranslator.com/widget/> [↑](#footnote-ref-71)
71. <http://support.microsoft.com/> [↑](#footnote-ref-72)
72. Metrika jeb koeficients automātiskai MT kvalitātes novērtēšanai (<http://en.wikipedia.org/wiki/BLEU>) [↑](#footnote-ref-73)
73. C. Callison-Burch, P. Koehn, C. Monz, J. Schroeder, Findings of the 2009 Workshop on Statistical Machine Translation, in Proceedings of the Fourth Workshop on Statistical Machine Translation, 1-28, Athens, Greece, 2009 [↑](#footnote-ref-74)
74. <http://www.statmt.org/moses/> [↑](#footnote-ref-75)
75. <http://translate.tilde.com> [↑](#footnote-ref-76)
76. <http://translate.google.com> [↑](#footnote-ref-77)
77. <http://www.microsofttranslator.com/> [↑](#footnote-ref-78)
78. <http://gridengine.sunsource.net/>, <http://en.wikipedia.org/wiki/Oracle_Grid_Engine> [↑](#footnote-ref-79)
79. http://www.likumi.lv/doc.php?id=154198&from=off [↑](#footnote-ref-80)
80. European Commission Machine Translation Service: <https://webgate.ec.europa.eu/mt/ecmt/> [↑](#footnote-ref-81)
81. <http://ixa.si.ehu.es/openmt2/> [↑](#footnote-ref-82)
82. <http://vertimas.vdu.lt/> [↑](#footnote-ref-83)
83. <http://www.epo.org/> [↑](#footnote-ref-84)
84. <http://tinyurl.com/epo-mt> [↑](#footnote-ref-85)
85. <http://www.lps.lv/> [↑](#footnote-ref-86)
86. <http://www.microsofttranslator.com/widget/> [↑](#footnote-ref-87)
87. Informācija no VA Kultūras informācijas sistēmas [↑](#footnote-ref-88)
88. http://www.valoda.lv/Valsts\_valoda/Latviesu\_valoda/mid\_502 [↑](#footnote-ref-89)
89. Centrālās statistikas biroja dati - IS06. VĪRIEŠU UN SIEVIEŠU VECUMA STRUKTŪRA GADA SĀKUMĀ [↑](#footnote-ref-91)