



Deliverable Task 6

ELRC Workshop Report for Latvia



Author(s):	Aivars Bērziņš (Tilde) Rihards Kalniņš (Tilde)
Dissemination Level:	Public
Version No.:	V1.0
Date:	02-11-2015



Contents

1	<u>ELRC Workshop in Latvia</u>	3
2	<u>Workshop Agenda</u>	4
3	<u>Summary of Content of Sessions</u>	7
3.1	Session 1: Opening addresses	7
3.2	Session 2: “The EU and Multilingualism”	7
3.3	Session 3: “The Latvian Language in the Digital Age”	8
3.4	Session 4: “Crossing the Language Barrier in Latvia’s Public e-Services”	8
3.5	Session 5: “Automated Translation: How does it work and what it is used for?”	9
3.6	Session 6: “Language Resources for Developing Automated Translation”	9
3.7	Session 7: “Language Resources in Latvia”	10
3.8	Session 8: “Getting Data and Language Resources: Technical and Practical Aspects”	10
3.9	Session 9: “Legal Framework for Contributing Data”	11
3.10	Session 10: “How can Public Institutions Benefit from Automated Translation Infrastructure?”	11
3.11	Session 11: “How Can I Contribute to CEF.AT: Practical Aspects”	12
3.12	Session 12: “Closing remarks”	12
4	<u>Workshop Presentation Materials</u>	13

1. ELRC Workshop in Latvia

This document reports on the ELRC Workshop in Latvia, which took place in Riga, on the 6th of October, at the office of the European Commission Representation in Latvia. Workshop was co-organized by ELRC Representative in Latvia – Tilde, Culture Information Systems Centre and Office of the European Commission Representation in Latvia.

Latvia ELRC workshop was attended by 59 participants covering a wide range of ministries and public organizations.

Event was video recorded and video are available on Workshop dedicated Web. Video are in Latvian and with voice over translations into English. Presentations are available on Workshop web site: http://lr-coordination.eu/riga_agenda.

After end of the workshop, follow up press release was issued and published on largest news agencies in Latvia LETA news portal.

Press release is available at: http://www.leta.lv/press_releases/74D1B573-8C22-4C4B-8056-A0193852C8F6/.

2. Workshop Agenda

**Agenda for CEF.AT - ELRC Training Workshop
(European Language Resource Coordination)
Office of the European Commission Representation in Latvia
Aspazijas bulvāris 28, Riga**

- 8.30 9.30 **Registration**
- 9.30 9.50 **Opening addresses**
Andrejs Vasiljevs, ELRC/Tilde
Andris Kužnieks, Representation of the European Commission in Latvia
Andrea Lösch, ELRC/DFKI
Saila Rinne, European Commission, Directorate-General Communications Networks, Content and Technology (DG CONNECT)
Armands Magone, Culture Information Systems Centre
- 9.50 10.05 **Multilingualism and Language Technologies for a United Europe**
Saila Rinne, European Commission, DG CONNECT
- 10.05 10.45 **Latvian Language in the Digital Age**
Panelists:
Andrejs Veisbergs, State Language Commission
Andrejs Vasiljevs, ELRC/Tilde
Normunds Grūzītis, University of Latvia, Institute of Mathematics and Computer Sciences
Inita Vītola, Latvian Language Agency
- 10.45 11.25 **Crossing the Language Barrier in Latvia's Public e-Services**
Panelists:
Gatis Ozols, Ministry of Environmental Protection and Regional Development
Armands Magone, Culture Information Systems Centre
Edgars Cīrulis, State Regional Development Agency
Anita Dūdiņa, Latvian Parliament
- 11.25 11.45 **Coffee break**

ELRC Workshop Report for Latvia

- 11.45 12.15 **Automated Translation: How Does it Work and What is it Used For?**
Inguna Skadiņa, ELRC/Tilde
- 12.15 12.40 **Language Resources for Developing Automated Translation**
Andrea Lösch, ELRC/DFKI
Andrejs Vasiljevs, ELRC/Tilde
- 12.40 13.40 **Lunch**
(at the conference venue)
- 13.40 14.20 **Language Resources in Latvia**
Panelists:
Jānis Ziediņš, Culture Information Systems Centre
Juris Baldunčiks, Latvian Academy of Science, Terminology Commission
Māris Baltiņš, State Language Centre
Ilze Auziņa, University of Latvia, Institute of Mathematics and Computer Science
Uldis Zariņš, Latvian National Library
- 14.20 14.40 **Getting Data and Language Resources: Technical and Practical Aspects**
Raivis Skadiņš, ELRC/Tilde
Eduards Cauna, Latvian Academy of Science, Terminology Commission
- 14.40 15.10 **Legal Framework for Contributing Data**
Prodromos Tsiavos, ELRC
- 15.10 15.30 **Coffee break**
- 15.30 16.10 **How can Public Institutions Benefit from Automated Translation Infrastructure?**
Saila Rinne, European Commission, DG CONNECT
Uldis Priede, European Commission, DG Translation
- 16.10 16.25 **How can I Contribute to CEF.AT: Practical Aspects**
Andrejs Vasiljevs, ELRC/Tilde
- 16.25 16.55 **How to Benefit from your Valuable Language Data and Technologies**

ELRC Workshop Report for Latvia

Open discussion

Saila Rinne, European Commission, DG CONNECT

Armands Magone, Culture Information Systems Centre

Andrejs Vasiljevs, ELRC/Tilde

16.55 17.00 **Wrap-up and Conclusions**

3. Summary of Content of Sessions

1. Session 1: Opening addresses

Andrejs Vasiljevs, the local ELRC representative, opened the event by welcoming the audience and introducing the key persons in conceiving and organizing the event, namely the ELRC consortium and the EC/DGT representatives.

Andris Kužnieks, Head of the EC Representation in Latvia, welcomed the audience to the European Union House and thanked the consortium for organizing the event on the premises. He expressed his interest in seeing the presentation and learning more about language technologies in Europe.

Saila Rinne, Programme Officer for EU Policies, DG Connect, introduced the audience to the European Commission's project for building CEF.AT and the role of language resources. She expressed her thanks to the local organizers for helping to organize the workshop.

Andrea Lösch, the ELRC representative in Germany, welcomed the audience on behalf of the consortium. She briefly described the importance of language resources for building CEF.AT, adding that she and her colleagues had recently organized the ELRC Workshop in Berlin, Germany.

Armands Magone, head of the Culture Information Systems Centre, also welcome the audience on behalf of his organization, one of the co-organizers of the event. Magone also briefly introduced the MT service Hugo.lv, which is maintained by the Culture Information Systems Centre and provides MT services for the Latvian public sector.

2. Session 2: “The EU and Multilingualism”

Saila Rinne, Programme Officer for EU Policies, DG Connect, began the first session following the opening addresses: “The EU and Multilingualism.” Ms. Rinne described the importance of linguistic diversity at the EU, home to 24 official EU languages and a total of about 60 major “regional/minority” languages. She stressed that the equality of the EU official languages is enshrined in the European legal basis, therefore the EU is committed to supporting multilingualism.

This was also the basis for the EU's strong initiative to support language technologies, such as machine translation. This support is exemplified in the creation of the MT@EC system for public administrations, as well as in the deployment of mature language technologies through the Connecting Europe Facility (CEF) Programme. EU support for multilingualism in Digital Europe has also resulted in the MT toolkit Moses, META-SHARE language resources, MT domain adapted pilots, standards and workflows for language processing, and tools for terminology work.

Rinne went on to describe the EU's Digital Single Market strategy and the ways in which language barriers affect public and private services. According to Rinne, pan-European public services currently face a multilingual challenge. There is no lingua franca among public administrations in Europe, and 90% of EU web users prefer to use their own language in online services. The traditional method used to counter this – human translation – is too expensive and too slow with the intended text volumes. Therefore, human translation is not a solution in all use scenarios.

The EU's solution to the challenge is to build a pan-European automated translation platform, CEF.AT, which will make European public online services multilingual by deploying mature language technologies (such as MT@EC) in a secure platform. The goals are to make public digital services equally usable by all EU users, irrespective of their working language and language skills, as well as to facilitate cross-border information exchange in public administration.

Rinne concluded by detailed the role of member states in the creation of CEF.AT, namely, the sharing of language resources as part of the ELRC project. By contributing resources, Rinne said, member

ELRC Workshop Report for Latvia

states could help improve the quality of the CEF.AT systems and then enjoy the benefits of automated translation: the reduction of language barriers in Europe.

3. Session 3: “The Latvian Language in the Digital Age”

This panel discussion focused on the Latvian language in the digital age – examining the status of “digital extinction” for Latvian and the technologies that are available to support the language today.

The panel consisted of four speakers: Andrejs Veisbergs, State Language Commission; Andrejs Vasiļjevs, ELRC/Tilde; Normunds Grūzītis, University of Latvia, Institute of Mathematics and Computer Sciences; and Inita Vītola, Latvian Language Agency.

In his presentation, Mr. Veisbergs spoke about the often voiced concern for the Latvian language’s “digital extinction.” Though he paid heed to the call and underscored its relevance, he also struck an optimistic chord, stressing that language technologies and the work of organizations like the State Language Commission, the Latvian Language Agency, and the Terminology Council were helping to support Latvian in the digital age.

For his part, Mr. Vasiļjevs shared the work of Tilde in developing various technologies for Latvian. These include proofing tools, terminology management services, machine translation systems, voice recognition software, among others. He drew particular attention to Tilde’s success in developing MT for Latvian, which has proven to provide significantly higher quality than Google Translate. He also briefly described Tilde’s work in developing the Latvian public sector’s MT service, Hugo.lv.

Mr. Grūzītis, who is a senior researcher at the University of Latvia’s Institute of Mathematics and Computer Sciences, discussed the cutting-edge research currently underway at the Institute. This includes research into morphological and syntactic analysis, semantic analysis, corpus linguistics, as well as speech recognition and synthesis. He also discussed the Institute’s work on developing data mining tools, which are used for media monitoring and website monitoring programmes. This work is going a long way toward supporting Latvian in the digital age.

Finally, Ms. Vītola of the Latvian Language Agency introduced the wealth of initiatives and activities undergone by her agency. This includes a series of educational workshops at Latvian universities called “The Latvian Language in the Digital Age,” the drafting of various policy documents about the Latvian language, the publication of online educational materials on the Latvian language (including various online games and fun language activities). Ms. Vītola also presented the online and off-line language materials specially developed for diaspora Latvians – which are going a long way toward helping keep the language alive outside of Latvia in the 21st century.

4. Session 4: “Crossing the Language Barrier in Latvia’s Public e-Services”

This session looked at the use of various language technology tools to cross language barriers in Latvia’s own e-services. Unlike many other countries, Latvia already has in place a MT service used by the public sector, making it a true success story in Multilingual Europe.

The panelists were: Gatis Ozols, Ministry of Environmental Protection and Regional Development; Armands Magone, Culture Information Systems Centre; Edgars Cīrulis, State Regional Development Agency; and Anita Dūdiņa, Latvian Parliament.

Mr. Ozols kicked off the session by introducing the wide range of e-services offered by the Latvian public sector. These e-services include cross-border digital signatures and national eID cards. He also cited the EU eGovernment Report from 2015, which states that 65% of respondents claimed that “language issues” were the biggest barrier keeping them from using public services across borders.

Mr. Magone continued the session by speaking about his organization’s own e-service: the MT service Hugo.lv. This service provides automated translation of texts, documents, and websites from English into Latvia, and vice versa, as well as from Latvian into Russian. The service is available as a

ELRC Workshop Report for Latvia

public website, at www.hugo.lv, and is integrated into various government portals and websites. This ensures that a wide range of residents have access to services and information in multiple languages.

Mr. Cīrulis, following up on the previous presentation about Hugo.lv, discussed the integration of the service into the Latvian government e-services portal, Latvija.lv. This portal, used by residents and citizens to receive public services and read relevant legislation and laws, includes a Hugo.lv widget. This guarantees that the texts can be instantly translated from Latvian into English (for international visitors) and Russian (for Latvia's sizeable Russian-speaking community).

Finally, Ms. Dūdiņa presented her work as head of the IT department of the Latvian parliament, or Saeima. The parliament has a substantial amount of translation work that needs to be performed, such as the translation of the website into English. In addition, parliament deputies often have the need to access legislative texts written in English. To help these deputies with translation – as not all of them speak English fluently – Dūdiņa has integrated Hugo.lv into the parliament's internal network. Using the MT tool, deputies can quickly translate English-language documents into Latvian, using them for general research purposes. Dūdiņa also foresees the further use of MT tools, in order to make the parliament's translation work more effective and productive.

5. Session 5: “Automated Translation: How does it work and what it is used for?”

Dr. Inguna Skadiņa gave a brief overview of the technical basics of automated translation. First she gave an overview of the history of MT, starting in the 1950s with the experiments at IBM. Then she elaborated on how early rule-based MT worked. Finally she introduced modern statistical machine translation, giving examples of statistical systems and how they are used in the decoding process. She talked about how sentence and word alignment functions, giving some interesting examples to help explain the complex points. Skadiņa also introduced the MOSES core, which forms the basis of various commercial MT platforms as well as MT@EC.

In conclusion, Dr. Skadiņa stressed the importance of language resources – monolingual data, bilingual data, glossaries, etc. – in building high-quality MT engines. She explained how much data was necessary for building systems, as well as the clear benefits of training domain-based MT systems.

6. Session 6: “Language Resources for Developing Automated Translation”

This session featured two speakers: Andrea Lösch, the ELRC representative in Germany, and Andrejs Vasiļjevs, the ELRC representative in Latvia.

Ms. Lösch focused on several questions: What is needed for MT? What counts as data for MT? What types of data can be used? She also discussed the difference between translations and aligned translation, as well as the role of dictionaries, ontologies, and terminologies. Ms. Lösch introduced the data format that are needed – both internet and digital data, and digital textual data. Finally, she discussed how language resources are produced from data, giving examples of how raw data become language resources.

In his half of the session, Mr. Vasiļjevs talked about how to extract language resources from data. He discussed various automated extraction methods, such as data extraction from online resources. He gave a key example of how data could be extracted from public sector websites – in his example, he showed a multilingual website of the Tourism Development Agency. Vasiļjevs also introduced the ILSP tool for gathering online data, which crawls online sites and links, extracts data, and aligns sentences. Vasiļjevs stressed, however, that public websites were only the “tip of the icebergs” in terms of online content. Infinitely more data is housed in organizations' internal systems. This is precisely the data and resources needed by the ELRC to help build the CEF.AT systems.

ELRC Workshop Report for Latvia

7. Session 7: “Language Resources in Latvia”

This large panel aimed at introducing the audience to the various language resources in Latvia, particularly those that had been accumulated and systematically compiled by various organizations.

Jānis Ziediņš of the Culture Information Systems Centre kicked off the discussion with his description of the Latvian language corpus that was used to build the Latvian MT service Hugo.lv. He described the efforts to gather hundreds of millions of parallel sentences in multiple languages – English, Russian, and Latvian. He also detailed the next steps for Hugo.lv – e.g., to gather cultural heritage data, metadata from libraries, and data in other specific domains. This will help make Hugo.lv more robust and applicable for use in other sectors, such as libraries, museums, and other memory institutions.

Next, the head of the Terminology Commission at the Latvian Academy of Sciences, Mr. Juris Baldunčiks, talked about the scope of the Commission’s activities. He described the ways in which the Commission manages and harmonizes terminology and ensures the consistent use of terminology. This is a particular challenge in Latvia, where terminologists must work to “localize” international terms, such as those related to IT and the web. The resources held by the Commission will be invaluable for making MT systems for Latvian more term-savvy.

The head of the Latvian Language Center, Mr. Māris Baltiņš, described the tasks of his organization. The Centre is tasked with protecting the rights and interests of the Latvian language, as well as ensuring the consistent use of Latvian in the public sphere. The centre also facilitates the continued use of Latvian in EU institutions and translates large volumes of official documentation, such as Latvian legislation and EU normative acts. The Centre also develops translation methods for these documents. Therefore the Centre holds invaluable language resources – many years of translation memories – that could potentially be collected and used to build CEF.AT.

Ms. Ilze Auziņa, of the University of Latvia’s Institute of Mathematics and Computer Science, also shared her organization’s experience with language resources. The Institute has been gathering language resources in Latvia for more than 25 years. These include electronic dictionaries and text and speech corpora. These resources have been used to build various language technology solutions, such as speech recognition and sentiment analysis tool, but could also be used as LRs to develop CEF.AT.

Finally, Mr. Uldis Zariņš of the Latvian National Library, presented the library’s digital cultural heritage data. This includes more than 4M books and periodicals in various languages, a text corpus, bibliographic databases, dictionaries, and other digital materials (such as maps and audio). These resources can be used to create parallel language resources. The library is also home to domain experts from various fields, who can provide input into the correct usage of the relevant language resources. Therefore the Latvian National Library is fully prepared to cooperate with the ELRC project.

8. Session 8: “Getting Data and Language Resources: Technical and Practical Aspects”

Representing the ELRC consortium, Mr. Raivis Skadiņš of Tilde described the data management workflow. He focused on several questions: What kind of data are we talking about? How does MT learn from data? Where can data be found?

Skadiņš described how data could be found online, in archives, and in the translation memories of localization providers. Then he introduced how data could be turned into language resources (i.e., the data management workflow). As a computational linguist, Skadiņš also discussed the technical aspects of gathering data – that is, the necessary file formats, alignment, data cleaning, data conversion, and data delivery. Skadiņš also provides insight into the fine points of data anonymization and licensing issues. Finally, Skadiņš talked about how, precisely, users could upload data to the ELRC repository and where to seek answers to questions.

ELRC Workshop Report for Latvia

The second half of the session consisted of a presentation by Eduards Cauna, who works for the Ministry of Regional Development and Environmental Protection. Cauna is also the head of the ICT sub-committee of the Terminology Commission at the Latvian Academy of Sciences. Therefore he was well-equipped to discuss the role of terminology in building language resources.

Mr. Cauna talked about the importance of terminology, where terms come from (i.e., scientific papers, public sector documentation, term dictionaries, educational materials, etc.), and where terms are compiled (i.e., various term banks and official term repositories). He also talked about the need to harmonize the use of terminology and merge various term banks into more easily accessible resources.

9. Session 9: “Legal Framework for Contributing Data”

In his presentation at the session on “Legal Framework for Contributing Data,” the ELRC legal expert Prodromos Tsiavos (The Media Institute, UCL) stressed the need for the following: a clear and easy ways to follow a regime for data re-use across the EU; legal and technical interoperability; and simple redress mechanisms. According to Tsiavos, the objective was to develop a single European market for innovative apps based on public data.

Tsiavos then posed an important question, probably on everybody’s minds: Does the PSI Directive make a difference? His answer was a resounding “yes”! According to Tsiavos, the Public Sector Information Directive (PSI) solves a number of issues: amends the existing PSI Directive; emphasizes open data; presents clearer cost rules; includes cultural Institutions (museums, libraries and archives) within the scope of the Directive; introduces a clear regime for exclusive agreements; and, finally, emphasizes standard and machine readable licenses.

Mr. Tsiavos then laid out the structure of rights, including PSI Rules, copyright rule, data protection rules, and excluded subject matter. Tsiavos then described how data would be used by ELRC, and indicated the five stages for releasing data: (1) exclude confidential information; (2) obtain prior informed consent, find a legal basis, anonymize or exclude personal data; (3) ensure there is no 3rd party copyrights, that the material is in the Public Domain or that the necessary licenses have been obtained; (4) follow the national PSI transposition rules (e.g. use the national Open Government licenses or the standard procedure for releasing PSI); (5) use a standard Open Government licence, open public license or reuse license, and follow the national or organisational PSI re-use policy.

In conclusion, Tsiavos presented eight compelling case studies of the best re-use info practices.

10. Session 10: “How can Public Institutions Benefit from Automated Translation Infrastructure?”

In this session, two representatives of the European Commission detailed how public institutions can benefit from automated translation infrastructure: Saila Rinne, of DG Connect, and Uldis Priede, of DG Translation.

Ms. Rinne discussed the importance of enabling multilingualism; the role of machine translation in doing so; the European Commission’s current MT service, MT@EC; and the future plans to build the CEF.AT platform. ELRC is working toward building CEF.AT. Rinne talked about the current interaction between actors in the member states and the EU, and the future vision, summed up as follows: “Wouldn’t it be great if I could start using a public service in any Member State from any place and obtain the information in my mother tongue?”

Rinne then discussed MT, which, she explained, is the only viable solution for the following: quick and cheap access to information in foreign languages; understanding information received in a foreign language that otherwise could not be used or would require substantial time and costs to translate; making multilingual use of websites possible; and facilitating cross-lingual information search and analytics.

ELRC Workshop Report for Latvia

According to Rinne, these benefits are apparent in the EC's current MT service, MT@EC, launched in 2013. MT@EC is a SMT system based on Moses. The system can be used to translate multiple documents in multiple languages, and is currently used by EU institutions and bodies.

In the future, Rinne said, MT@EC will evolve into CEF.AT. CEF.AT will build on the existing MT@EC service - but not be limited to it; put emphasis on secure, quality, customisable MT for pan-European online services, but not be limited to them; and be a multilingualism enabler – not only MT.

For his part, Uldis Priede, of DG Translation, talked about how translators currently work at the EU. Translators are responsible for translating content into all official EU languages. More than 2 000 translators are currently employed by DG Translation, while another 5 000 work at EU institutions. In 2014, they translated more than 2.3M pages.

According to Priede, MT is substantially helping to make this process more productive at the EU level. This will only continue in the future. His main messages were that MT is a tool that must be learned in order to continue working in the translation industry; and if the data for MT is good, then the MT system will be good as well.

11. Session 11: “How Can I Contribute to CEF.AT: Practical Aspects”

In this final session, Andrejs Vasiljevs of ELRC/Tilde began by posing several questions that were probably on the minds of listeners: Does my organization have data that can contribute to the CEF.AT infrastructure? Can I make this data available? What are the practical and organizational aspects? Can AT benefit my organization? How can I follow the development of the AT infrastructure?

Vasiljevs then attempted to answer the questions in turn. He started by once again showing the cycle of language resource management, stressing that organization should always ask for TMs from their translation providers. He then detailed the work of the ELRC consortium, describing how the consortium works, who is involved, and what are the goals and scope of the project.

Vasiljevs then turned the attention of listeners to the ways in which they could get involved in the project, as well as once again underscored the national anchor point in Latvia: the Culture Information Systems Centre.

12. Session 12: “Closing remarks”

The closing session consisted of brief closing remarks from the main organizers of the event: Andrejs Vasiljevs of ELRC/Tilde, Saila Rinne of DG Connect, and Jānis Ziediņš of the Culture Information Systems Centre.

Each of the speakers gave his or her concluding remarks and, at the request of the moderator, shared the lessons learned throughout the date. The speakers once again stressed the need for collaboration not only between EU Member States, but also by individual institutions within each country. Only by collaborating and sharing resources can the vision of CEF.AT be finally realized, reducing language barriers in Europe.

4. Workshop Presentation Materials

All presentations are provided on ELRC Riga Workshop Web site (http://lr-coordination.eu/riga_agenda). This Web site contains all presentations in .pdf format and are supplemented with video recordings of presentations. Two types of video are provided – in original language and voice over translations into English for presentation originally in Latvian and voice over in Latvian for presentations in English.

To provide more information about workshop topic, ELRC Riga Workshop dedicated Brochure was prepared. Brochure is adjusted to the style of the ELRC web page. Brochure provides information about multilingualism, CEF.AT and CEF activities, ELRC (activities and consortium), what are important resources. As well information about activities in Latvia on Automated translation solution for Public administration. Brochure is provided further and is available through ELRC Web site - <http://www.lr-coordination.eu/latvia>.

Eiropas valodu resursu koordinācijas aktivitātes Latvijas semināra dienas kārtība

Laiks	Tēma
8.30 - 9.30	Reģistrācija
	Atklāšanas uzrunas (Video - ENG, LAT)
	Andrejs Vasiljevs, ELRC/Tilde
	Andris Kužnieks, Eiropas Komisijas pārstāvniecība Latvijā
9.30 - 9.50	Andrea Lösch ELRC/ Vācijas Mākslīgā intelekta pētniecības centrs (DFKI)
	Saila Rinne, ES politikas programmu vadītāja, Eiropas Komisijas Komunikācijas tīklu, saturs un tehnoloģiju ģenerāldirektorāts (DG CONNECT)
	Armands Magone, Kultūras informācijas sistēmu centra direktors
	Daudzvalodība un valodas tehnoloģijas vienotākai Eiropai (Video - ENG, LAT)
9.50 - 10.05	Saila Rinne, ES politikas programmu vadītāja, Eiropas Komisijas Komunikācijas tīklu, saturs un tehnoloģiju ģenerāldirektorāts (DG CONNECT) (Prezentācija - ENG, LAT)

Picture 1 Snapshot of Workshop Web page (http://lr-coordination.eu/riga_agenda)



Valodu daudzveidība – Eiropas stūrakmens

Eiropas Valodas resursu koordinācijas seminārs

Rīga, Latvija
2015. gada 6. oktobrī



Figure 1 Front cover of Brochure



Attēls: Eiropas Komisija



Ceļā uz vienotu daudzvalodu Eiropu

Pēdējo gadu laikā Eiropa ir veiksmīgi novērsusi daudzus ierobežojumus, kas kavē vienoto digitālo tirgu. Tā ir viena no Eiropas galvenajām prioritātēm. Nauda, preces un cilvēki tagad var brīvi ceļot pār valstu robežām, bagātinot Eiropas Savienību kopumā.

Tomēr vēl joprojām pastāv vairāki ierobežojumi, kas kavē vienota digitālā tirgus izveidi un iespēju visiem iedzīvotājiem gūt labumu no veiksmīgāk integrētas Eiropas. Viens no pēdējiem šķēršļiem ir valodas barjera.

Kaut gan valodu daudzveidība ir ES stūrakmens, daudzvalodība rada arī barjeras starp tautām. Valodas barjeras liedz Eiropas patērētājiem pilnvērtīgi izmantot tiešsaistes iespējas un ierobežo piekļuvi daudziem digitālajiem pakalpojumiem.

Šo barjeru dēļ ES tiešsaistes tirgus ir sadrumstalots. 43% eiropiešu nekad neiegādājas tiešsaistē pieejamās preces un pakalpojumus, ja tie tiek piedāvāti svešvalodā; valodas robežas ietekmē piekļuvi valsts nodrošinātajiem e-pakalpojumiem; un ES bagātais kultūras saturs nevar nonākt ārpus valodas kopienu robežām.

Risinājums, kas var pārvarēt valodas barjeras digitālajā vidē, ir valodas tehnoloģija.

Tehnoloģijas sasniegumi, piemēram, automātiskā tulkošana, paver plašas daudzvalodu saziņas iespējas digitālajā vidē. Automātiskā tulkošana veicina saikni starp iedzīvotājiem, valsts iestādēm un nevalstiskajām organizācijām, panākot ciešāku komunikāciju, iesaistīšanos un sapratni.

Tā rodas ne vien patiesi vienots daudzvalodu digitālais tirgus, bet arī satuvināta, veiksmīgāk integrēta daudzvalodu Eiropa.



Picture 2 Page 1 and page about multilingualism in Europe



Eiropas automatizētās tulkošanas ietvars CEF.AT

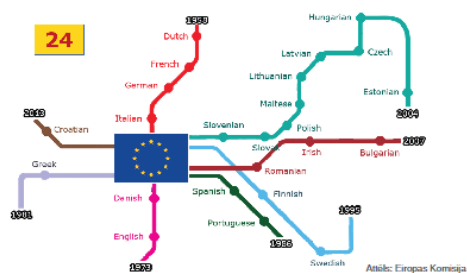
Eiropas infrastruktūru savienošanas programma (Connecting Europe Facilities – CEF) finansē projektus, kas nodrošina trūkstošo saikni starp Eiropas transporta, enerģētikas un digitālo tīklu pamatstrukturām. Tāpat tā padara Eiropas ekonomiku zaļāku, veicinot videi draudzīgākus transporta veidus, ātrgaitas plaīstas pieslēgumus un sekmējot atjaunojamo enerģoresursu izmantošanu saskaņā ar stratēģiju "Eiropa 2020". Koncentrējoties uz savstarpēji saistītiem, vieniem un ilgtspējīgiem transporta, enerģētikas un digitālajiem tīkliem, Eiropas infrastruktūras savienošanas instruments sekmēs āgtspējīga Eiropas vienotā digitālā tīrgu izveidi.

Eiropas infrastruktūru savienošanas programmā būtiska loma atvēlēta **telekomunikācijas un IKT sektoram**. Telekomunikācijas un IKT jomā paredzēta tīklu un digitālo pakalpojumu infrastruktūru (**Digital Service Infrastructure - DSI**) izveide. Tās nodrošinās plašu pārrobežu pakalpojumu tīklu izveidi, uzņēmiem un valsts pārvaldes iestādēm. Viena no nozīmīgākajām ir visas ES dalībvalstīs aptveroša virtuāla infrastruktūra automatizētas tulkošanas nodrošināšanai (**CEF.AT**).

CEF.AT vizija ir vienotā platformā piedāvāt **augstas kvalitātes automatizētas tulkošanas sistēmas** visām Eiropas Savienības oficiālajām valodām dažādās specializētās jomās. CEF.AT platformā tiks ņemti vērā dažādi valsts pārvaldes scenāriji patērētāju tiesību, veselības aizsardzības, publisko iepirkumu, sociālās aizsardzības, kultūras un citās jomās.

CEF.AT vēlas ietvert labākos risinājumus un tehnoloģijas, kas radītas ES un tās dalībvalstīs, lai radītu unikālu un **aptverošu risinājumu valodas barjeru pārvēršanai** un vienotākas Eiropas izveidei.

Daudzvalodu pieeja veicinās Eiropas valstu tiešsaistes pakalpojumu izmantošanu. Izmantojot automatizētas tulkošanas iespējas, valsts iestādes visā Eiropā spers soli tuvāk darbam **bez valodas barjerām**.



ELRC - Eiropas Valodu resursu koordinācijas platforma

Ar Eiropas Valodu resursu koordinācijas platformas (ELRC) izveidi Eiropas Komisija sāk līdz šim neparedzētu valodas datu apkopošanas darbu. Tas ir pirmais solis ceļā uz CEF.AT pielāgošanu valsts nodrošināto pakalpojumu vajadzībām visās ES dalībvalstīs, kā arī Islandē un Norvēģijā, tā veicinot daudzvalodu pakalpojumu sniegšanu Eiropas iedzīvotājiem, pārvaldes iestādēm un uzņēmumiem. Tādējādi ELRC ne vien likvidēs plaisu starp pašreizējo mašintulkošanas sistēmu iespējām, kuras Eiropas Komisija piedāvā valsts pārvaldes iestādēm, un visu Eiropas valstu nodrošināto pakalpojumu faktiskajām ikdienas vajadzībām, bet arī nepastarpināti atbalstīs Eiropas tautu valodas pašos lietošanas pamatos.

ELRC mērķis ir izveidot pastāvīgu valodas resursu koordinēšanas mehānismu, lai nodrošinātu Eiropas automatizētas tulkošanas ietvaram nepieciešamos valodas resursus. Tikai ar katras ES dalībvalsts aktīvu iesaisti **valodas resursu** apkopošanā un **mašintulkošanas tehnoloģiju** attīstībā ir iespējams radīt līdzvērtīgas tulkošanas iespējas visām ES valodām. ELRC veicina nacionālās aktivitātes šai jomā un to rezultātu pielietojumu kopējas Eiropas infrastruktūras izveidei.

ELRC darbība aptver visas **24 oficiālās ES valodas, kā arī norvēģu un islandiešu valodas**. Padarot pieejamus apjomīgus un daudzveidīgus valodas resursus, ELRC palīdz uzlabot automatizētas tulkošanas kvalitāti, daudzpusību un piemērotību dažādiem risinājumiem. Īpašu vērību CEF pievērš vajadzībām, ko nosaka nepieciešamība padarīt CEF digitālos pakalpojumus daudzvalodīgus.

Šā mērķa īstenošanai ELRC sevisku uzmanību pievērš valodas resursiem, kas tiek radīti un glabāti publiskās pārvaldes iestādēs. Šajās iestādēs top visdažādākie dokumenti, daudzi no kuriem tiek tulkti citās valodās. Tie var kalpot par vērtīgu materiālu **automatizētas tulkošanas sistēmu uzlabošanai**. Arī akadēmiskajā vidē pētniecības un mācību procesā top visdažādākie valodas resursi.

ELRC rūpējas, lai šie valodas dati tiktu apzināti un padarīti pieejami gan automatizētas tulkošanas uzlabošanai, gan pēc iespējas arī citiem pielietojumiem, sekojot atvērto datu principam.

ELRC palīdz publiskā sektora pārstāvjiem un citiem valodas resursu turētājiem uzlabot šo datu pārvaldību, risināt ar to pieejamību un izmantošanu saistītos tehniskos un juridiskos jautājumus. Noderīgi padomi un atbalsta centra kontaktinformācija sniegta ELRC mājaslapā: www.lr-coordination.eu.

Reizi gadā ELRC rīko **Eiropas mēroga valodas konferences**, kas pulcē Eiropas Komisijas pārstāvjus un dalībniekus no ES valstu publiskā sektora, pētniecības institūcijām un tehnoloģiju uzņēmumiem. Pirmā šāda konference notika 2015. gada aprīlī Rīgas Samitā par daudzvalodu vienoto digitālo tīrgu ietvaros. Konferences materiāli pieejami **Rīgas Samitā** un ELRC mājaslapās.

Picture 3 CEF.AT and ELRC activities



Hugo.lv - Latvijas piensums Eiropas automatizētās tulkošanas ietvaram

Hugo.lv ir Latvijas valsts pārvaldes mašintulkošanas pakalpojums, kas brīvi pieejams jebkuram Latvijas iedzīvotājam. Tas nodrošina automatizētu tulkošanu latviešu-angļu valodā un pretēji, kā arī latviešu-kirovu valodā. Lietotāji var tulkot tekstu, dokumentus un interneta lapas.

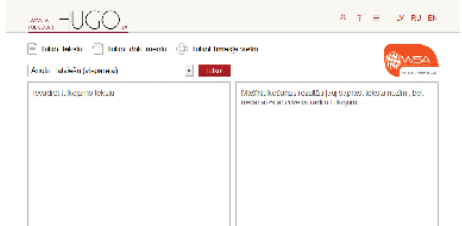
Hugo.lv ir īpaši pielāgots latviešu valodai un valsts pārvaldes dokumentiem, tāpēc tulkošanas kvalitāte ir daudz augstāka, nekā tulkojot ar citiem tiešsaistes tulkošanas pakalpojumiem.


Priekšrocības:

- Piemērots teksta, dokumentu un interneta lapu tulkošanai
- Īpaši izstrādāts latviešu valodas vajadzībām, ietverot valodai specifiskus moduljus
- Īpaši pielāgots valsts pārvaldes dokumentiem
- Tulkošana ir droša, un dati netiek nodoti trešajām pusēm


Hugo.lv var integrēt visās valsts pārvaldes platformās un mājaslapās, veicinot daudzvalodu pieeju e-pārvaldes informācijai un e-pakalpojumiem. Pakalpojums ir integrēts e-pārvaldības platformā Latvija.lv, paverot piekļuves iespējas portāla e-pakalpojumiem un informācijai lietotājiem no visas pasaules.


Hugo.lv projektā tika izveidots pasaules lielākais latviešu valodas korpus. Tajā ir vairāk nekā 100 miljoni teikumu no ļoti daudzveidīgiem avotiem, piemēram: valsts pārvaldes dokumenti, publiski pieejamie resursi, terminu saraksti, drukāti teksti un atsevišķi elektroniski izdevumi no interneta. Tika veikta tīmekļa pārlūkošana un saturs izgūšana no nozarei tuvu mājaslapām.






IEGULDĪJUMS TAVĀ NĀKOTNĒ







Hugo.lv

Latvijas valsts sektora iestādēm paredzētais inovatīvais pakalpojums Hugo.lv izstrādāts ar Eiropas Savienības līdzfinansējumu, sadarbojoties Kulturalis informācijas sistēmu centram (KISC) un valodas tehnoloģiju uzņēmumam Tilde kopīgā projektā „Daudzvalodu korpusa un mašintulkošanas infrastruktūras izveide e-pakalpojumu pieejamības nodrošināšanai”.




Lai saņemtu informāciju par Hugo.lv izmantošanas iespējām valsts pārvaldes iestāžu darbā un e-pakalpojumos, lūdzam, sazinieties ar lapas uzturētājiem (Kulturalis informācijas sistēmu centrs).

E-pasts: help@kis.gov.lv

Picture 4 Latvia Public administration MT solution Hugo.lv



Valodas resursi automātiskai tulkošanai

Lai uzlabotu CEF automātiskās tulkošanas platformas pakalpojumu kvalitāti, tulkošanas sistēmām nepieciešami apjomīgi valodas resursi (teksti, vārdnīcas, t.sk. terminoloģiskās vārdnīcas, ģeogrāfisko u.c. nosaukumu, apzīmējumu, saīsinājumu u.c. datubāzes, u. tml.) visās to valstu oficiālajās valodās, kas piedalās CEF programmā.

Dalībvalstu un ar CEF saistītās valsts pārvaldes, nevalstiskās un privātās organizācijas katru dienu rada lielu vērtīgu valodas datu apjomu. Vairākums šo datu ir nepieciešami CEF-AT platformas pilnveidei.

Valodas resursi ir nepieciešami, lai uzlabotu gan vispārējām vajadzībām, gan konkrētām nozarēm domātās mašintulkošanas kvalitāti. Vajadzīgie resursi ietver gan lielus paralelos (piemēram, likumdošanas u. c. dokumentus un to tulkojumus) un monolingvalos korpusus (piemēram, Nacionālos korpusus, ko daudzas ES dalībvalstis veido savām valodām). Specializētu mašintulkošanas sistēmu izstrādei lieti noder ne tikai vispārīgi resursi, bet arī specifiski konkrētas nozares valodas dati – šai jomai raksturīgi teksti, to tulkojumi, termini, klasifikatori u. c., īpaši tādi, kas saistīti ar publiskiem tiesiskajiem pakalpojumiem.

Tekstu korpusi - apjomā liels daudzveidīgu tekstu kopums, kas parasti uzkrāts elektroniski un saistīts ar programmatūru, kura atvieglo tā lingvistisko analīzi (Valodniecības pamatterminu skaidrojošā vārdnīca. – R., 2007). Tekstu korpusu veido, apkopojot lielu daudzumu atsevišķu dokumentu, teksta vienību, tīmekļa lapu saturu u. tml.

Paralēlais korpusi - tekstu korpusi divās valodās (avotvalodā un mērķvalodā), kur katram avotvalodas tekstam ir atbilstošs tulkojums mērķvalodā. Piemēram, Latvijas Republikas tiesību akti latviešu valodā un to tulkojumi angļu valodā. Mūsdienu automatizētās tulkošanas sistēmas ar statistisku metožu palīdzību no paralēlajiem korpusiem spēj "iemācīties" tulkošanas sakarības.

Terminoloģijas vārdnīcas - nozaru specifiskas vārdnīcas, kas satur terminus un to tulkojumus divās valodās (avotvalodā un mērķvalodā). Terminoloģijas vārdnīcas ļauj pielāgot automatizētās tulkošanas sistēmas konkrētām nozarēm, nodrošinot, ka konkrēto nozaru tekstu tulkojumos terminoloģija tiek lietota konsekventi un pareizi.





ELRC konsorcijs

Eiropas Komisija



Vācijas Mākslīgā intelekta pētniecības centrs



DFKI ir pakļautas pētniecības iestādes Kaiserslauternē, Zārbrikenē un Brēmenē, kā arī projektu birojs Berlīnē. DFKI ir vadošais pētniecības centrs Vācijā novatorisku komerciālas programmatūras tehnoloģiju nozarē, kas izmanto mākslīgo intelektu.

Tilde



Tilde ir valodas tehnoloģijas uzņēmums, kura specializācija ir pielāgotu mašintulkošanas sistēmu terminoloģijas mācīšanas pakalpojumu izstrāde. Uzņēmums dibināts 1991. gadā, un tam ir filiāles Rīgā, Tallinā un Viņņā.

Novērtējumu un valodas resursu izplatīšanas aģentūra



ELDA darbojas Parīzē (Francija), un tās mērķis ir veicināt valodas resursu izmantošanu cilvēka valodas tehnoloģijas (Human Language Technology – HLT) sektorā un novērtēt valodas inženierijas tehnoloģiju.

Valodas un runas apstrādes institūts



Valodas un runas apstrādes institūts ir pētniecības institūts, kas darbojas Grieķijas Izglītības un reliģisko lietu ministrijas Pētniecības un tehnoloģiju ģenerālsekretariāta paspārnē.

Tulkošanas automatizācijas lietotāju sabiedrība



TAUS ir globālo valodu un tulkošanas industrijas resursu centrs, kura galvenais birojs atrodas Amsterdamā. Tā mērķis ir panākt augstāku tulkošanas kvalitāti, izmantojot novatorisku pieeju un automatizāciju.

Picture 5 Language resources and ELRC consortium

 Pasākuma programma		 Pasākuma programma	
8:30 – 9:30	Reģistrācija	13:40 – 14:20	Valodas resursi Latvijā Diskusijas dalībnieki: Jānis Ziediņš, Kultūras informācijas sistēmu centrs Raivis Skadiņš, ELRC/Tilde Juris Baldunčiks, Latvijas Zinātņu akadēmijas Terminoloģijas komisija Māris Baltiņš, Valsts Valodas centrs Ilze Auziņa, Latvijas Universitātes Matemātikas un informātikas institūts Uldis Zariņš, Latvijas Nacionālā bibliotēka
9:30 – 9:50	Atkliāšanas uzrunas Andrejs Vasiljevs, ELRC/Tilde Andris Kužnieks, Eiropas Komisijas pārstāvniecības Latvijā vadītājas vietnieks Jozefs van Genabits, ELRC/Vācijas Mākslīgā intelekta pētniecības centrs Sailla Rinne, ES politikas programmu vadītāja, Eiropas Komisijas Komunikācijas tīklu, saturs un tehnoloģiju ģenerāldirektorāts (DG CONNECT) Armands Magone, Kultūras informācijas sistēmu centra direktors	14:20 – 14:40	Valodaa resursu ieguve: praktiskie un tehniskie aspekti Raivis Skadiņš, ELRC/Tilde Eduards Cauna, Latvijas Zinātņu akadēmijas Terminoloģijas komisija
9:50 – 10:05	Daudzvalodība un valodaa tehnoloģijas vienotākaai Eiropai Sailla Rinne, Eiropas Komisija, DG CONNECT	14:40 – 15:10	Tiesiskā bāze datu izmantošanas veicināšanai Prodromos Tsiavos, ELRC
10:05 – 10:45	Latviešu valoda digitālajā vidē Diskusijas dalībnieki: Andrejs Veisbergs, Valsts valodaa komisija Andrejs Vasiljevs, ELRC/Tilde Normunds Grūziņis, Latvijas Universitātes Matemātikas un informātikas institūts Inīta Vitola, Latviešu valodaa aģentūra	15:10 – 15:30	Kafijas pauze
10:45 – 11:25	Valodaa barjeru mazināšana Latvijaa publikajaa e-pakalpojuma Diskusijas dalībnieki: Gatis Ozols, Vides aizsardzības un reģionālās attīstības ministrija Armands Magone, Kultūras informācijas sistēmu centrs Edgars Ciniņš, Valsts reģionālās attīstības aģentūra Anita Dūdiņa, Latvijas Republikas Saeima	15:30 – 16:10	Kādu labumu no automātiskāa tulkošanas infrastruktūras var gūt valsts iestādes Sailla Rinne, Eiropas Komisija, DG CONNECT Uldis Priede, Eiropas Komisija, Tulkošanas ģenerāldirektorāts
11:25 – 11:45	Kafijas pauze	16:10 – 16:25	Kā varu iesaistīties automātiskāa tulkošanas infrastruktūras attīstībā Andrejs Vasiljevs, ELRC/Tilde
11:45 – 12:15	Automātiskāa tulkošana: kā tā darbojas, un kam to izmanto Jozefs van Genabits, ELRC/Vācijas Mākslīgā intelekta pētniecības centrs	16:25 – 16:55	Kā gūt maksimālo labumu no valodaa datiem un tehnoloģijām Atvērta diskusija
12:15 – 12:40	Valodaa dati automātiskāa tulkošanaa attīstībai Inguna Skadiņa, ELRC/Tilde	16:55 – 17:00	Noelģuma aecinājumi
12:40 – 13:40	Puadīnaa (semināra telpāa)		



Atsāk: Eiropas Komisija

Picture 6 Agenda

Organizatori



Latvijas koordinators



Mājas lapa: www.lr-coordination.eu/
E-pasts: info@lr-coordination.eu

Picture 7 Back cover