

# Collecting Parallel Data for Clients and Building Customized NMT



SMART-Select Workshop #4 on Data Curation for (Neural) Machine Translation

Luxembourg, 20 November 2018

Raivis Skadiņš, Tilde, Director of Research and Development

# Outline

- Highlights from different data collection projects
- Challenges & Solutions
- What and Why We do in a Different Way
- Tools
- Customized NMT engines





# Some of our recent Data Collection Projects

- Estonian Open Parallel Corpus
- Tilde Model Corpus, the ODINE Open Data Incubator
- Presidency of the Council of the European Union
- European Language Resource Coordination (ELRC)
- Latvian e-government MT Platform
- Business Clients



# Estonian Open Parallel Corpus & Tilde Model (ODINE)

- There is public funding for open [language] data
  - National Programme for Estonian Language Technology
  - ODINE Open Data Incubator for Europe
  - etc.
- Let's collect data we need and share it with others
- Estonia: ET, EN, RU, FI, any but preferably wide domain
- ODINE: all EU languages, finances, medicine, tourism





# Estonian Open Parallel Corpus & Tilde Model (ODINE)

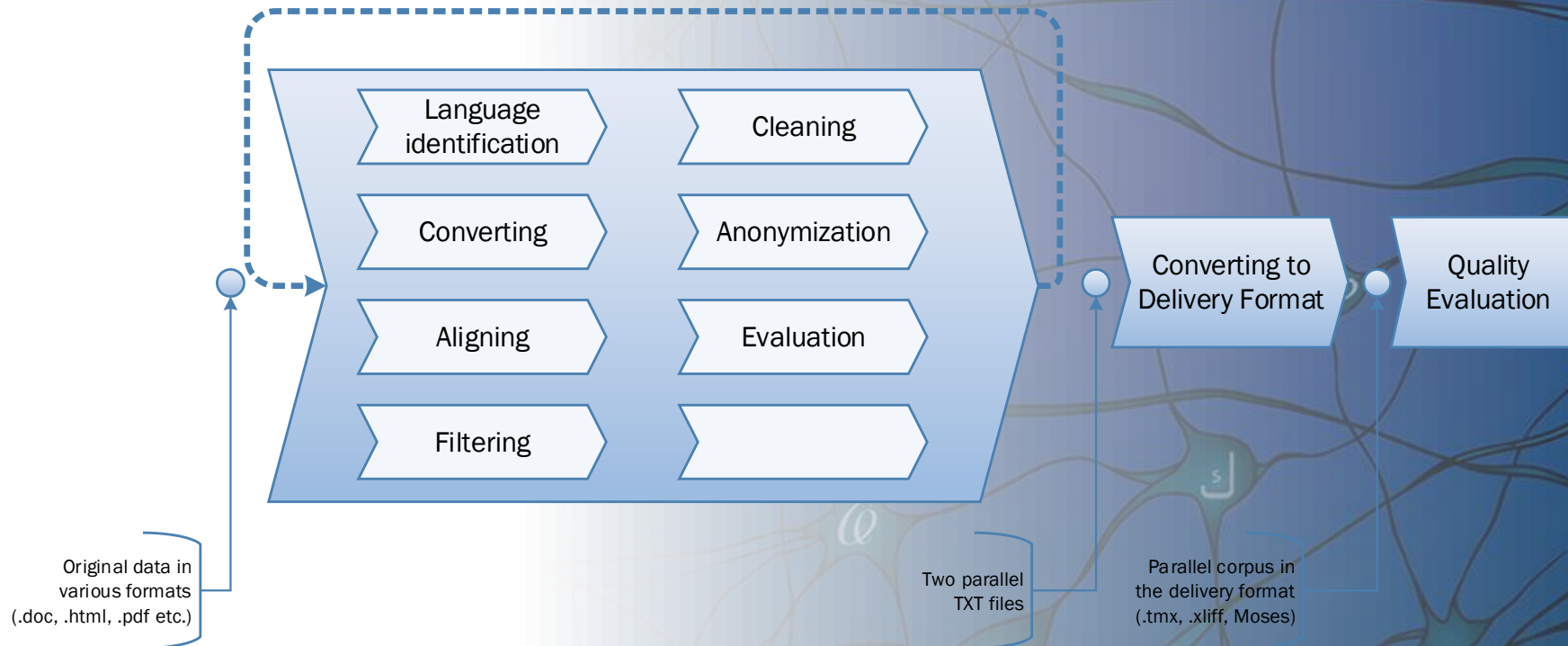
- Challenges

- A lot of small websites
- There is almost no ET-RU parallel content
- Comparable, localized, adapted, transcreated content
- “deep web”, regular crawlers do not crawl it
- DOC, PDF, WordPerfect



# General workflows

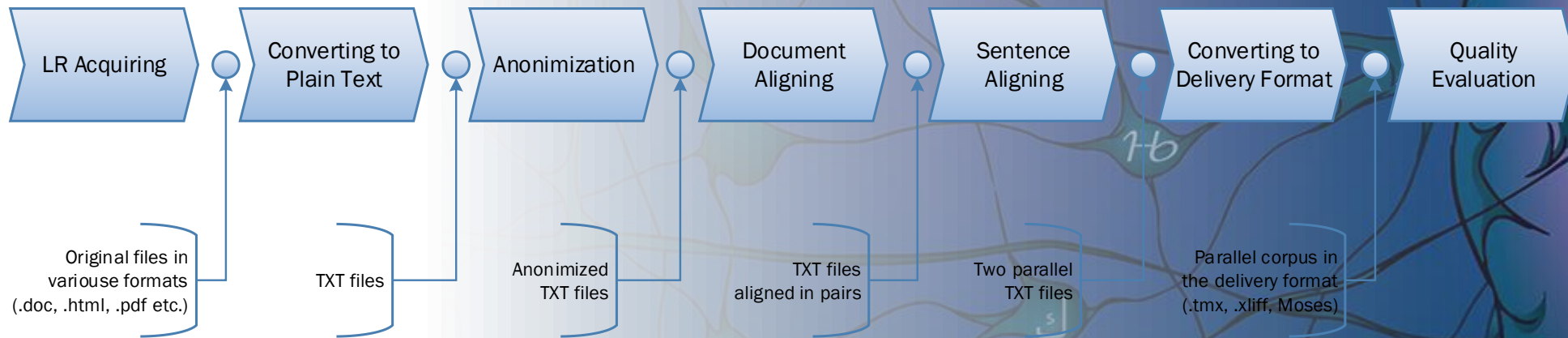
## General language resource processing workflow





# General workflows

## General parallel corpus processing workflow



# Data Processing Workflows

- We aim at high quality
- Manual website inspection
  - Content quality: professional vs amateur, translated vs comparable or transcreated or even MT
  - Structure: URL-s, alignments, achieves, site maps, search options
  - Size estimation





# Data Processing Workflows

- Document alignment
  - It's always better to get it from the site structure, document IDs or links
  - If not, use any hints – pictures, authors, dates
  - Alignment tools
- Sentence alignment
  - We use Microsoft Sentence Alignment of Bilingual Corpora (Moore, 2002)
  - It's robust, it tolerates comparable documents, **but** we had to rewrite it 😊
- Filtering, automatic corrections

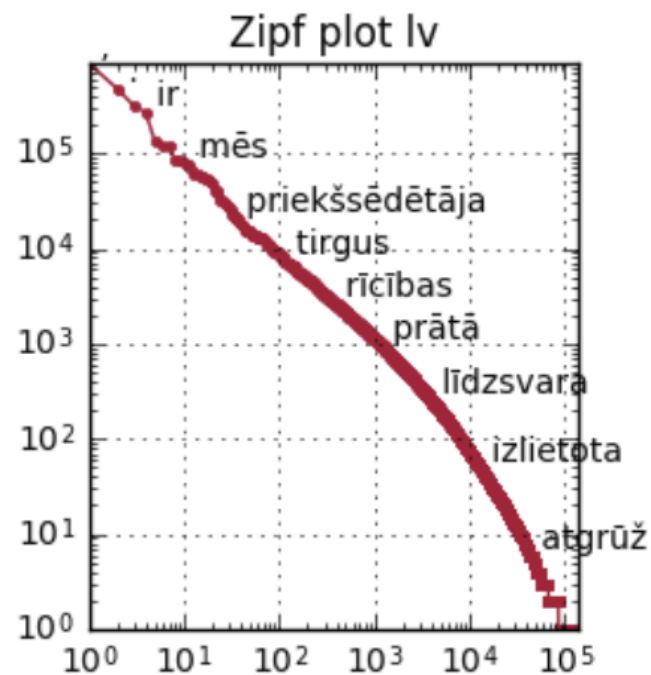
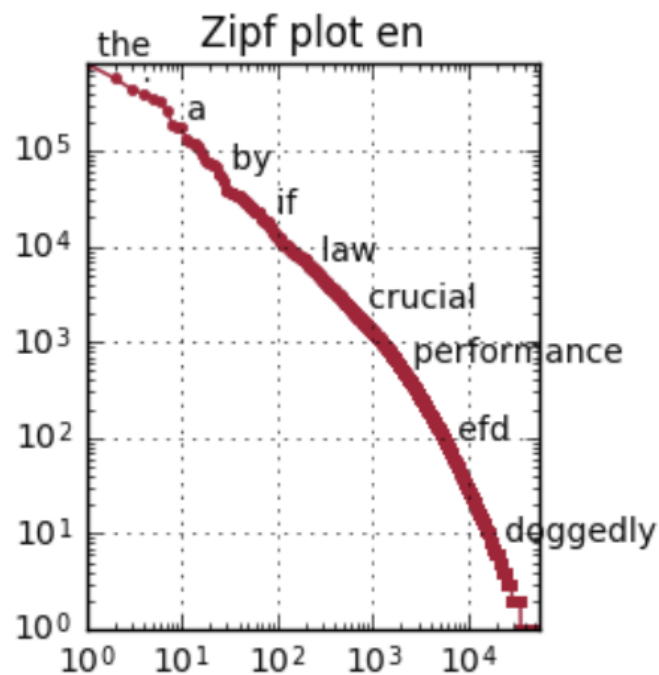


# Data Processing Workflows

- Evaluation

- Automatic: spelling errors, length ratio, Zipf plots
- Manual

Measurement / language	en	lv	en-lv
segments	453,504	453,504	453,504
words	12,048,961	9,919,756	
spelling-errors	165,608	160,582	
Sentences removed by filter			
EqualLinesParallelFilter	2,345	2,345	2,345
LineLengthRatioParallelFilter	175	175	175
UniqueLineFilter	15,186	15,186	15,186
MaxLineLengthFilter	14	14	14
ForeignWordsFilter	11,505	11,505	11,505
MaxWordLengthFilter	2	2	2
UniqueParallelLineFilter	11,403	11,403	11,403
Sentences normalized by cleaner			
RemoveTags	38	38	38





# Tilde MODEL Corpus

- European Economic and Social Committee documents
- RAPID - Press Release Database of EC
- European Central Bank
- European Medicines Agency
- World Bank
- AirBaltic, LiveRiga and other tourism sites
- It's public, CC-BY 4.0

Tilde MODEL Corpus

Language pairs: 69

	bg	cs	da	de	el	en	es	et	fi	fr	hr	hu	is	it	lt	lv	mt	nl
de	1.2M	1.4M	2.4M		2.4M	2.9M	2.4M	1.3M	2.2M	2.8M	226K	1.2M	209	2.4M	1.3M	1.3M	1.2M	2.4M
en	1.2M	1.3M	2.4M	2.9M	2.4M		2.5M	1.3M	2.1M	3.1M	250K	1.2M	220	2.6M	1.4M	1.4M	1.2M	2.4M
fr	1.2M	1.4M	2.4M	2.8M	2.5M	3.1M	2.6M	1.3M	2.1M		249K	1.2M	207	2.6M	1.4M	1.4M	1.2M	2.4M

**Tilde MODEL - RAPID**

Tilde MODEL - RAPID multilingual parallel corpus is compiled from all press releases of Press Release Database of European Commission released between 1975 and end of 2016 as available from <http://europa.eu/rapid/>.

Languages: 25  
Language pairs: 72

	bg	cs	da	de	el	en	es	et	fi	fr	hr	hu	is	it	lt	lv	mt	nl
de	177K	209K	499K		484K	1.0M	529K	188K	373K	983K	41K	194K	179	528K	177K	187K	186K	52
en	199K	243K	537K	1.0M	558K		684K	227K	435K	1.5M	50K	236K	195	673K	213K	233K	231K	62
fr	195K	234K	513K	983K	545K	1.5M	677K	208K	397K		49K	230K	211	661K	211K	223K	222K	60

**Tilde MODEL - ECB**

Tilde MODEL - ECB multilingual parallel corpus is compiled from the multilingual pages of European Central Bank web site <http://ebc.europa.eu/>.

Languages: 23  
Language pairs: 232

	cs	da	de	el	en	es	et	fi	fr	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl
bg	3K	3K	3K	455	3K	3K	3K	2K	4K	2K	3K	3K	3K	3K	4K	4K	3K	4K	4K	3K	4K
cs		4K	3K	4K	3K	4K	3K	3K	4K	2K	4K	3K	4K	4K	3K	4K	4K	4K	3K	4K	4K
da			4K	4K	4K	5K	3K	4K	5K	2K	4K	5K	4K	4K	4K	5K	4K	5K	3K	4K	3K
de				4K	4K	4K	3K	4K	4K	1K	2K	4K	3K	3K	3K	5K	3K	4K	2K	3K	3K
el					5K	6K	4K	4K	6K	2K	4K	6K	4K	4K	5K	6K	4K	6K	4K	5K	5K
en						5K	3K	4K	5K	1K	3K	5K	3K	4K	4K	5K	3K	5K	3K	4K	4K
es							4K	5K	7K	2K	4K	7K	4K	4K	5K	7K	4K	6K	4K	5K	5K
et								4K	4K	2K	4K	4K	4K	4K	4K	4K	4K	4K	3K	4K	4K
fi									4K	1K	3K	4K	3K	3K	3K	5K	3K	4K	2K	3K	3K
fr										2K	4K	6K	4K	5K	5K	7K	4K	7K	4K	5K	5K
hr											2K	2K	2K	2K	2K	2K	2K	2K	2K	2K	2K
hu												4K	4K	4K	4K	4K	4K	4K	4K	4K	4K
it													4K	4K	5K	7K	4K	6K	4K	4K	5K

# European Language Resource Coordination (ELRC)





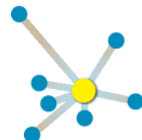
# CEF eTranslation & European Language Resource Coordination



<http://www.lr-coordination.eu/>



European  
Commission



European Language  
Resource Coordination  
*Connecting Europe Facility*

# ELRC: Challenges

- Language Resource Coordination ☺
- Getting data for EC eTranslation
  - All EU languages
  - Cleared IPR





# ELRC: Solutions

- Network of National Anchor Points (NAP)
- Awareness workshops
- Data donations
  - Lithuanian Parliament, Estonian President, Statoil etc.
- Data collection
  - Websites (ELRC consortium and NAPs)
    - Governments, municipalities, agencies, museums
    - What works in one state, should work in others
  - Crawling & Processing (ELRC)
- Onsite assistance



# ELRC: IPR

- Public Sector Information
- Licensing Contracts
- No *grey* or *dark* content

## Intellectual property (Copyright)

This website, all the elements contained in it (including the layout) and the information and Services are protected by the relevant intellectual property and copyright legislation.

Unless otherwise stated, the Luxembourg State grants no license or authorization with regard to the intellectual property rights which it holds in respect of this site, the elements contained in it or the Services. Moreover, reproduction of the information or Services, either wholly or in part and in whatever form or by whatever means, is not permitted without the prior written consent of the Ministries responsible for this site.

Unless otherwise stated, users are authorized to consult, download and print the available documents and information, on the following conditions:

- documents may only be used for personal purposes, for information and in a strictly private context;
- documents and information may not be modified in any way whatsoever;
- documents and information may not be disseminated outside or beyond this website.

The rights impliedly or expressly granted to you above constitute an authorization to use the guichet.lu website; in no circumstances do they constitute a transfer or assignment of property rights or other rights in relation to this site.





# ELRC: IPR

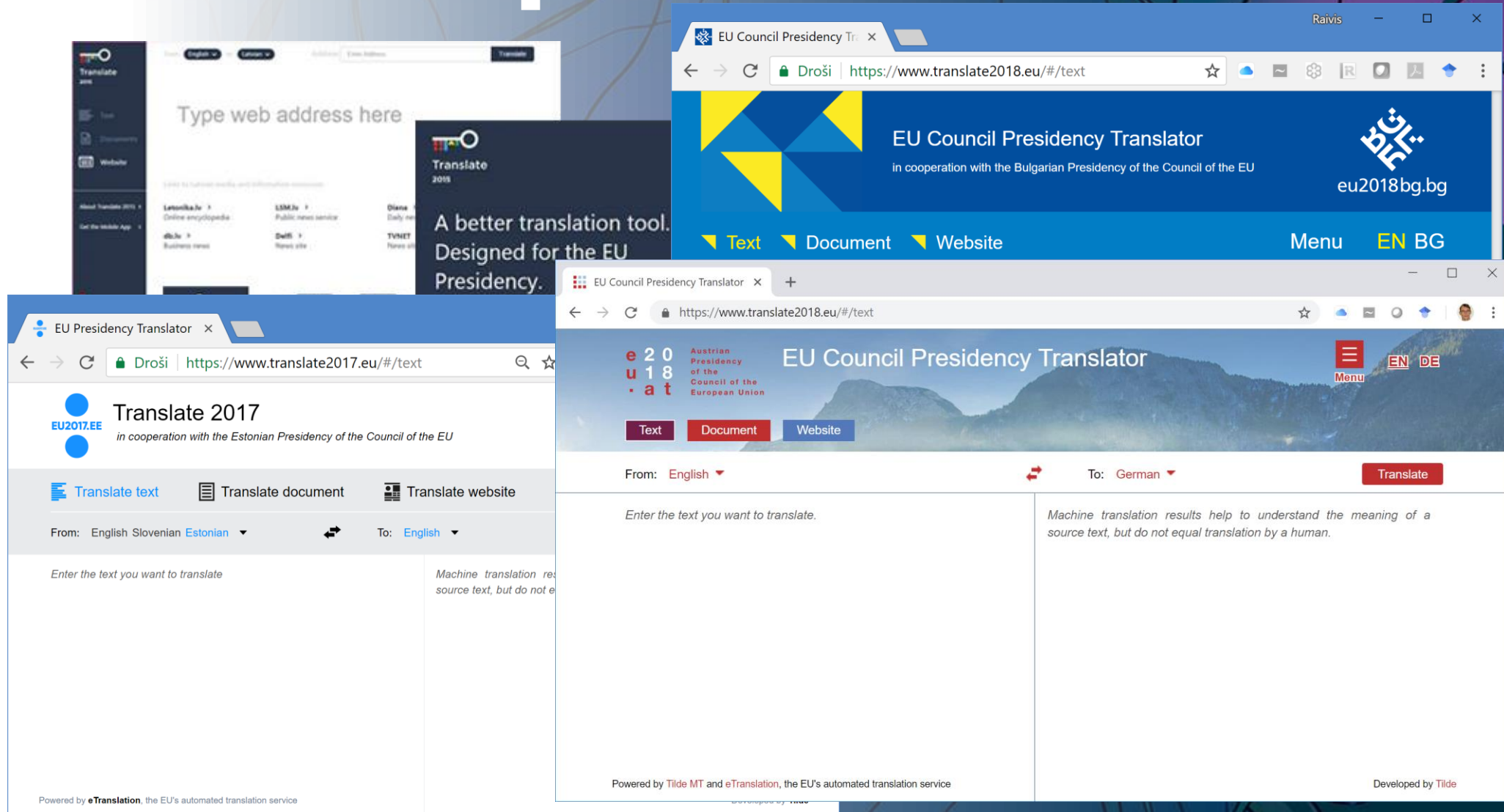
- Public Sector Information
- Licensing Contracts
- No *grey* or *dark* content



The screenshot shows a web browser window with the URL [www.eu2015lu.eu/en/sup...](http://www.eu2015lu.eu/en/sup...). The page header includes the Luxembourg Government logo and the text "LE GOUVERNEMENT DU GRAND-DUCHÉ DE LUXEMBOURG". Below this, the "GRAND DUCHY OF Luxembourg" logo is displayed, along with the text "Presidency of the Council of the European Union". A blue navigation bar contains a menu icon, a search icon, and language selection buttons for "fr", "de", and "en" (which is highlighted). The main content area is titled "Legal aspects" and includes social media icons for Facebook, Twitter, Email, and Print. The text on the page states: "This site is provided by the **Ministry of Foreign and European Affairs** and the **Information and Press Office of the Government**. Any person using the information, documents, products, software and/or services(hereinafter collectively referred to as the "**Services**") offered by this website shall be deemed to be aware of, and to have accepted, all the provisions of these general terms and conditions of use."

# MT for Presidencies of the Council of the European Union

- Latvia
- Estonia
- Bulgaria
- Austria



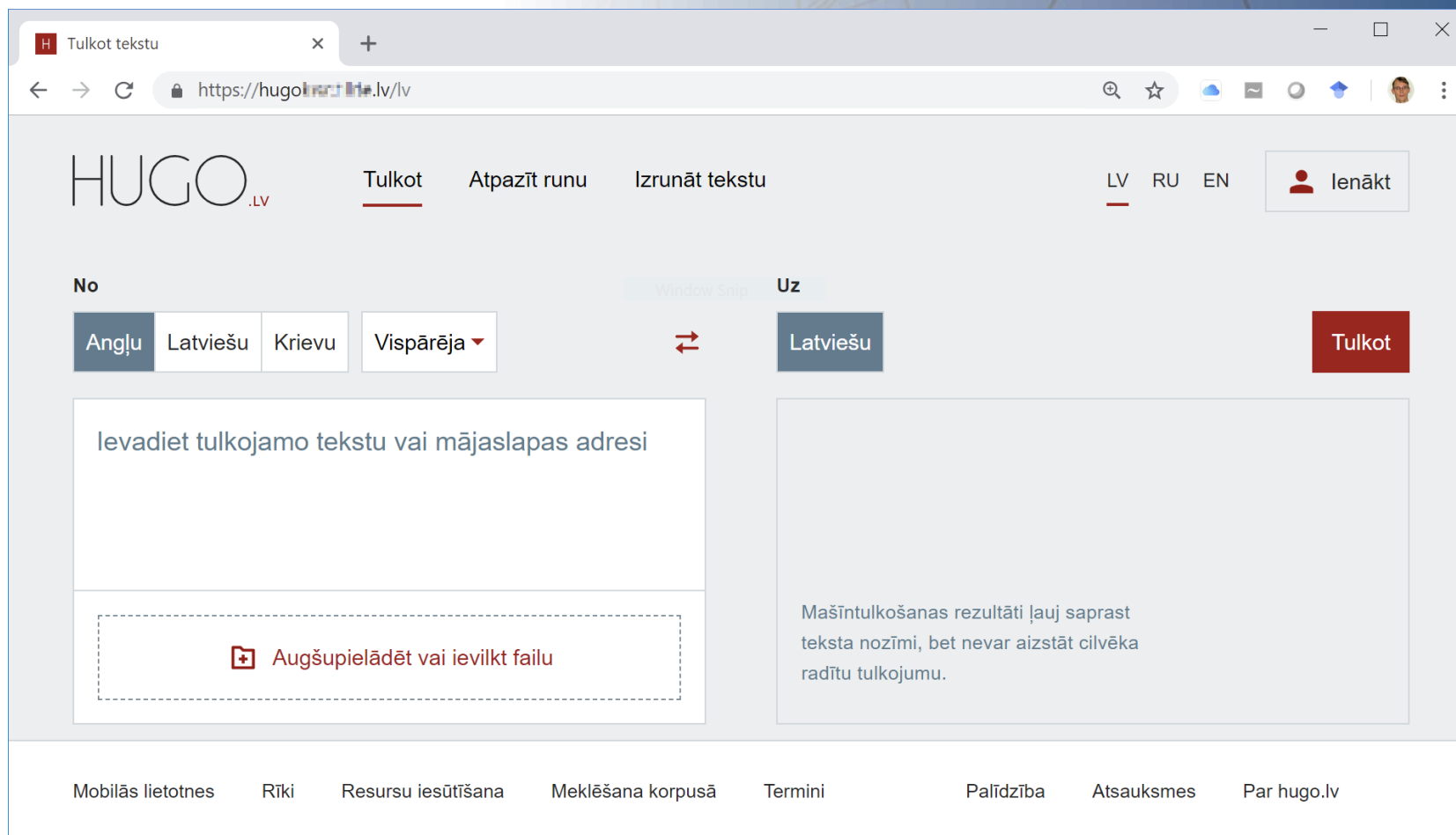


# Challenges & Solutions

- Political News Domain
- Training data
- Local Partners
  - Help with data
  - Visibility



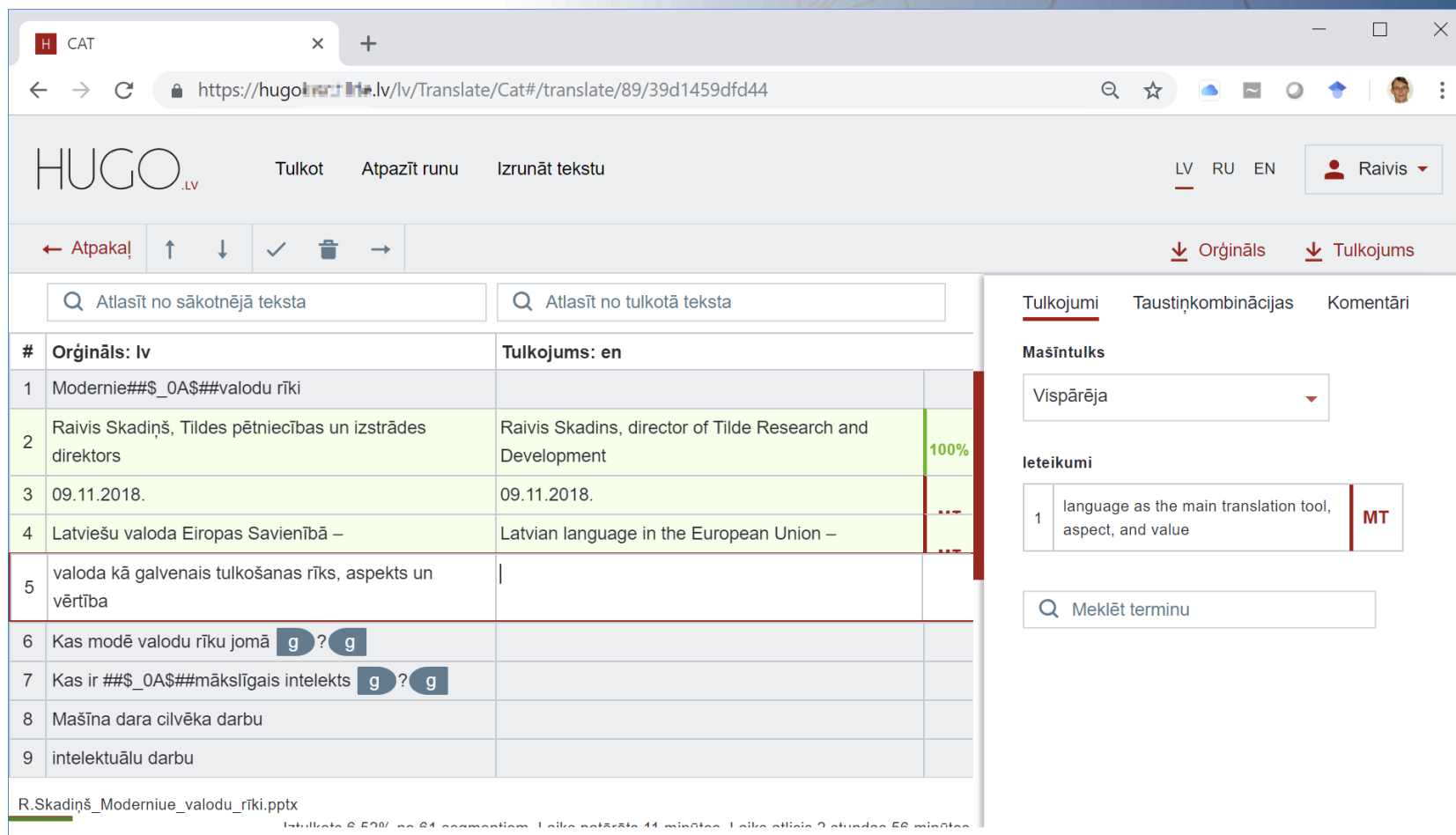
# HUGO.lv: NMT for Latvian Public Administration





# HUGO.lv: NMT for Latvian Public Administration

CAT for data collection; MT + TM



HUGO.lv

Tulkot Atpazīt runu Izrunāt tekstu

LV RU EN Raivis

← Atpakaļ ↑ ↓ ✓ ☒ →

Orģināls Tulkojums

Atlasīt no sākotnējā teksta Atlasīt no tulkotā teksta

#	Orģināls: lv	Tulkojums: en	
1	Modernie##\$_OAS##valodu rīki		
2	Raivis Skadiņš, Tildes pētniecības un izstrādes direktors	Raivis Skadins, director of Tilde Research and Development	100%
3	09.11.2018.	09.11.2018.	
4	Latviešu valoda Eiropas Savienībā –	Latvian language in the European Union –	
5	valoda kā galvenais tulkošanas rīks, aspekts un vērtība		
6	Kas modē valodu rīku jomā <b>g</b> ? <b>g</b>		
7	Kas ir ##\$_OAS##mākslīgais intelekts <b>g</b> ? <b>g</b>		
8	Mašīna dara cilvēka darbu		
9	intelektuālu darbu		

R.Skadiņš\_Modernie\_valodu\_rīki.pptx

Izskatās 6.52% no 64 segmentiem. Laiks patērēts 11 minūtes. Laiks atlikis 2 stundas 58 minūtes.

Tulkojumi Taustiņkombinācijas Komentāri

Mašīntulks

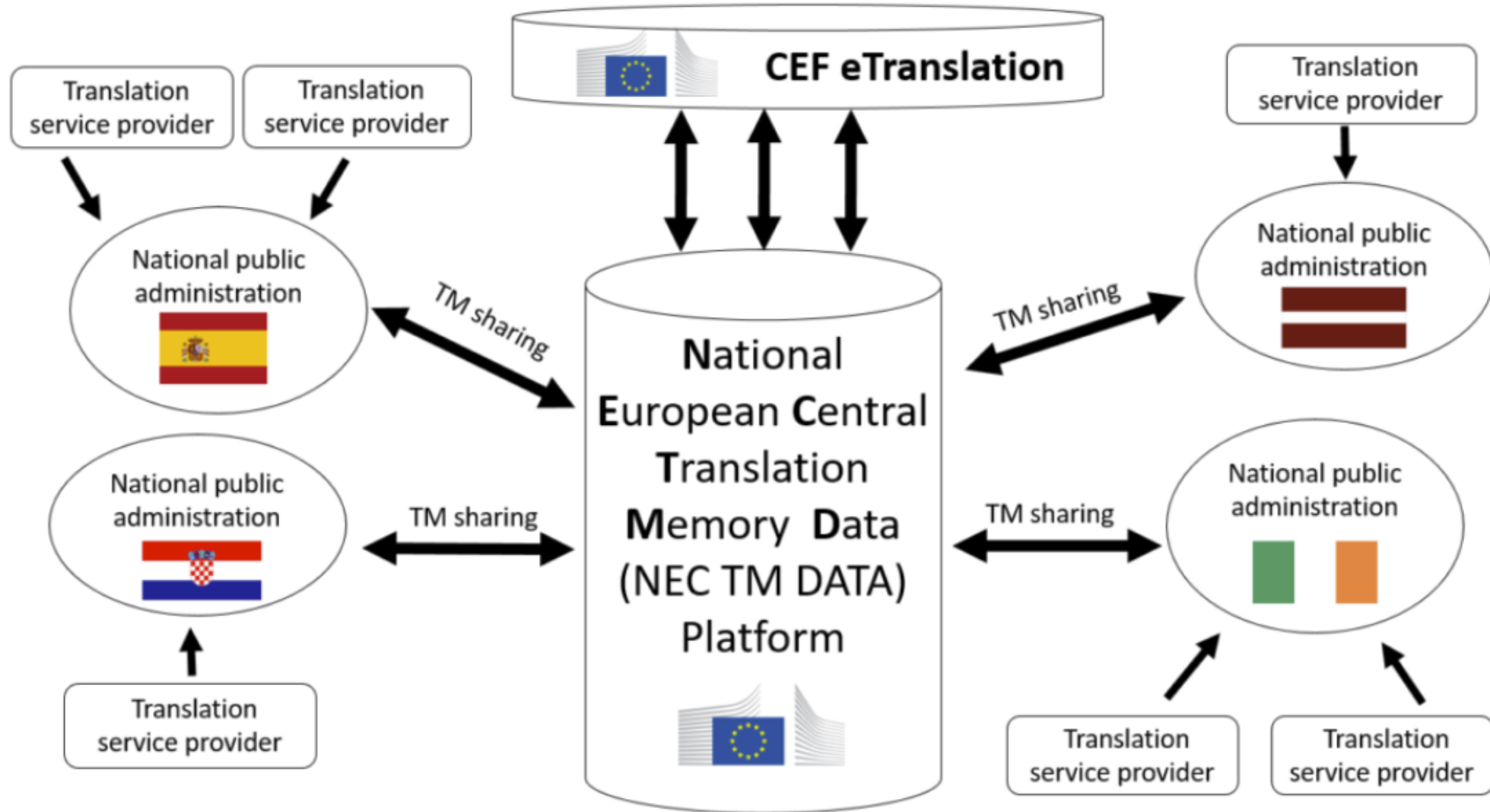
Vispārēja

Ieteikumi

1 language as the main translation tool, aspect, and value MT

Meklēt terminu







# Challenges & Solutions

- Data Collection 😊
- Work with public sector institutions
  - State language center
  - Court administration
- Content licensing
  - Rīgas laiks, Kodex+



# Business Clients

- Ideal case: TMX, XLIFF-s, Trados TM-s
- Reality:
  - few translated Word or even PDF documents
  - Excel with dozen for terms
  - Sometimes a lot of monolingual data





# Tools

- Crawling
  - Bitextor
  - ILSP focused crawler
  - Teleport Ultra
  - Custom created PERL and Python scripts
- Conversion PDF-(HTML)-Plaintext
  - Adobe Acrobat Pro
  - Microsoft Word



# Customized NMT Engines

- We start with a general domain NMT engine
- Training with in-domain data
  - Parallel
  - Back-translated (in several iterations)

