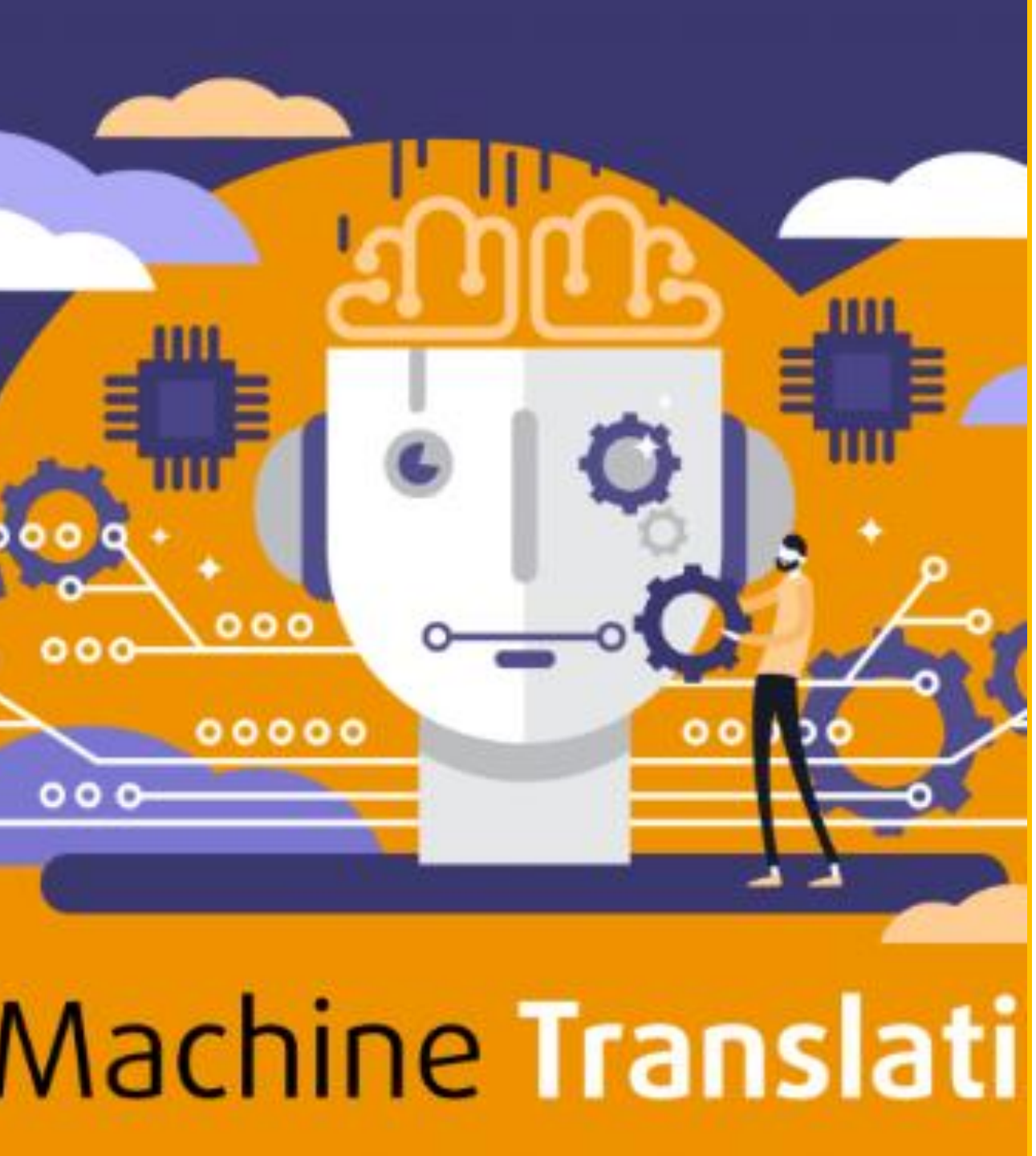
Masintõlke areng Euroopa Komisjonis: eesti ja teiste väikekeelte kitsaskohad ja lahendused

Kristiina Suviste
Euroopa Komisjoni kirjaliku tõlke peadirektoraadi informaatikaosakond



Täna käsitletavad teemad

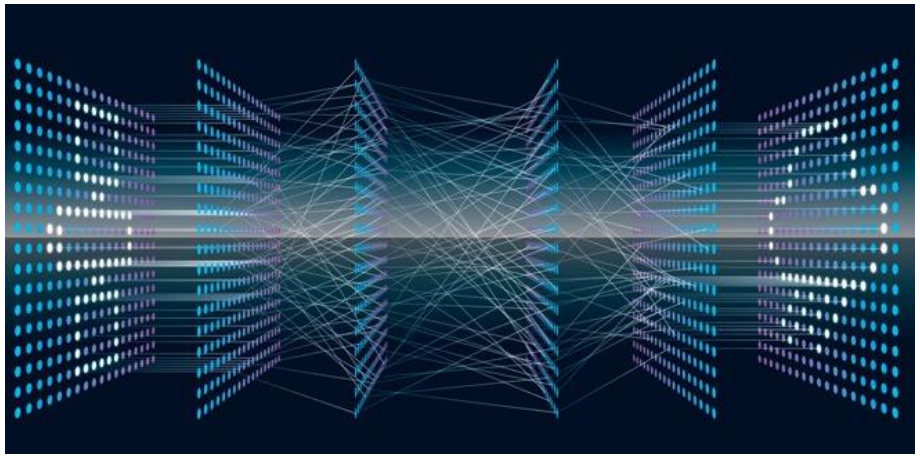
- Kuidas masin keelt „näeb“ ja tõlgib
- Masintõlke tööpõhimõtted
- Masintõlke väljakutsed üldisemalt
- Masintõlke väljakutsed väikekeelte kontekstis: eesti keele näitel
- Masintõlke areng Euroopa Komisjonis
- Kokkuvõte ja küsimused

European Commission

Kuidas masin keelt näeb?

Kuidas masin keelt „näeb“ ja tõlgib

- Teksti teisendamine digitaalsesse vormingusse
- Mustrid ja struktuur (loomuliku keele töötlemine – NLP)
- Mudelite/mootorite treenimine
- Vastuste genereerimine



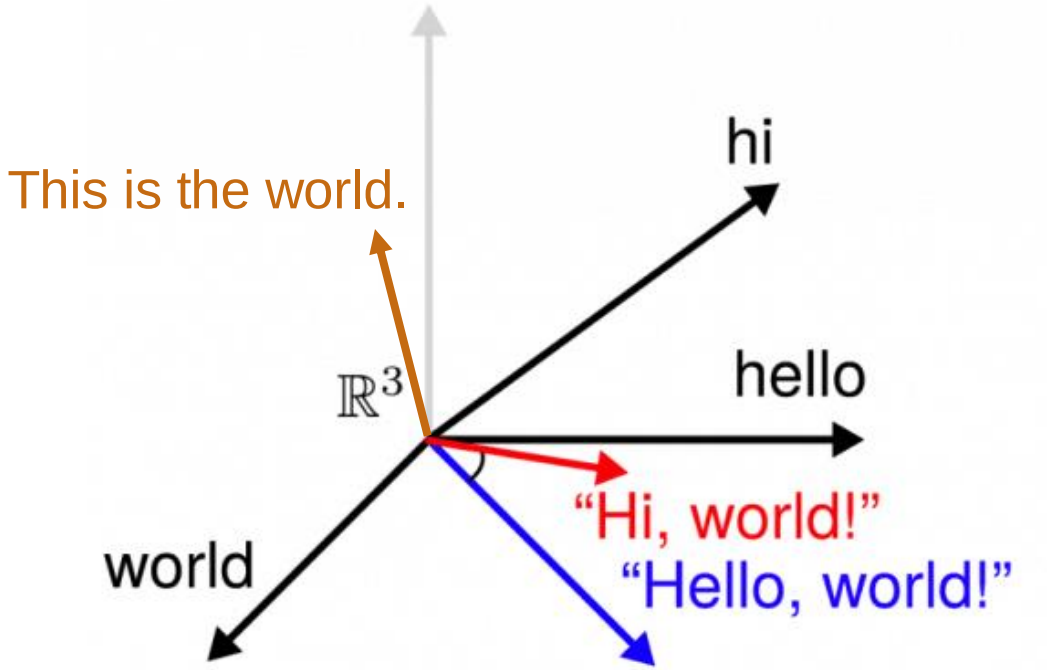
European Commission

Teksti teisendamine digitaalsesse vormingusse

Transformermudelid

Tekst muudetakse mõõtmepunktideks ja vektoriteks

```
 1.3286e-01, -1.2067e-01, -4.4166e-01, -1.9887e-01,  1.7223e-01,
-3.8480e-02,  3.8753e-01, -4.6423e-01,  2.3012e-01, -2.5965e-01,
 1.8702e-01, -2.6960e-01,  1.6820e-01, -4.1661e-01, -1.6216e-01,
-1.2017e-01,  6.6695e-02,  2.3583e-01, -2.7710e-01, -2.2654e-01,
-2.5241e-02, -4.6770e-01, -2.5526e-01,  1.8228e-01, -1.0045e-01,
-1.9519e-01, -4.5371e-02, -1.0781e-01, -3.6034e-02,  3.8383e-02,
 7.2391e-01,  3.7758e-01, -3.3185e-01,  6.9592e-02,  2.4604e-01,
-2.3778e-01, -1.2194e-02, -9.7566e-03,  3.4999e-02,  3.2956e-02,
 8.1280e-02, -3.5268e-01, -8.1572e-02,  3.7758e-01,  1.1375e-01,
 3.6395e-01, -2.4066e-01,  3.7662e-01, -1.0668e-01,  2.3552e-01,
 2.4936e-01,  7.6259e-03, -2.8949e-01,  2.2092e-02,  9.3504e-02,
```



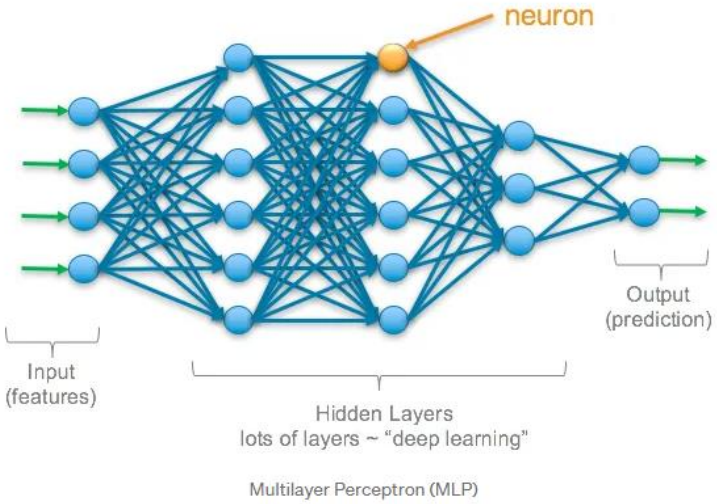
Source: https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/soft_cosine_tutorial.ipynb

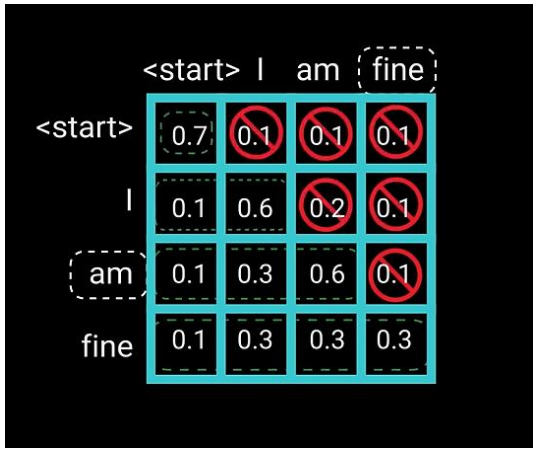
European Commission

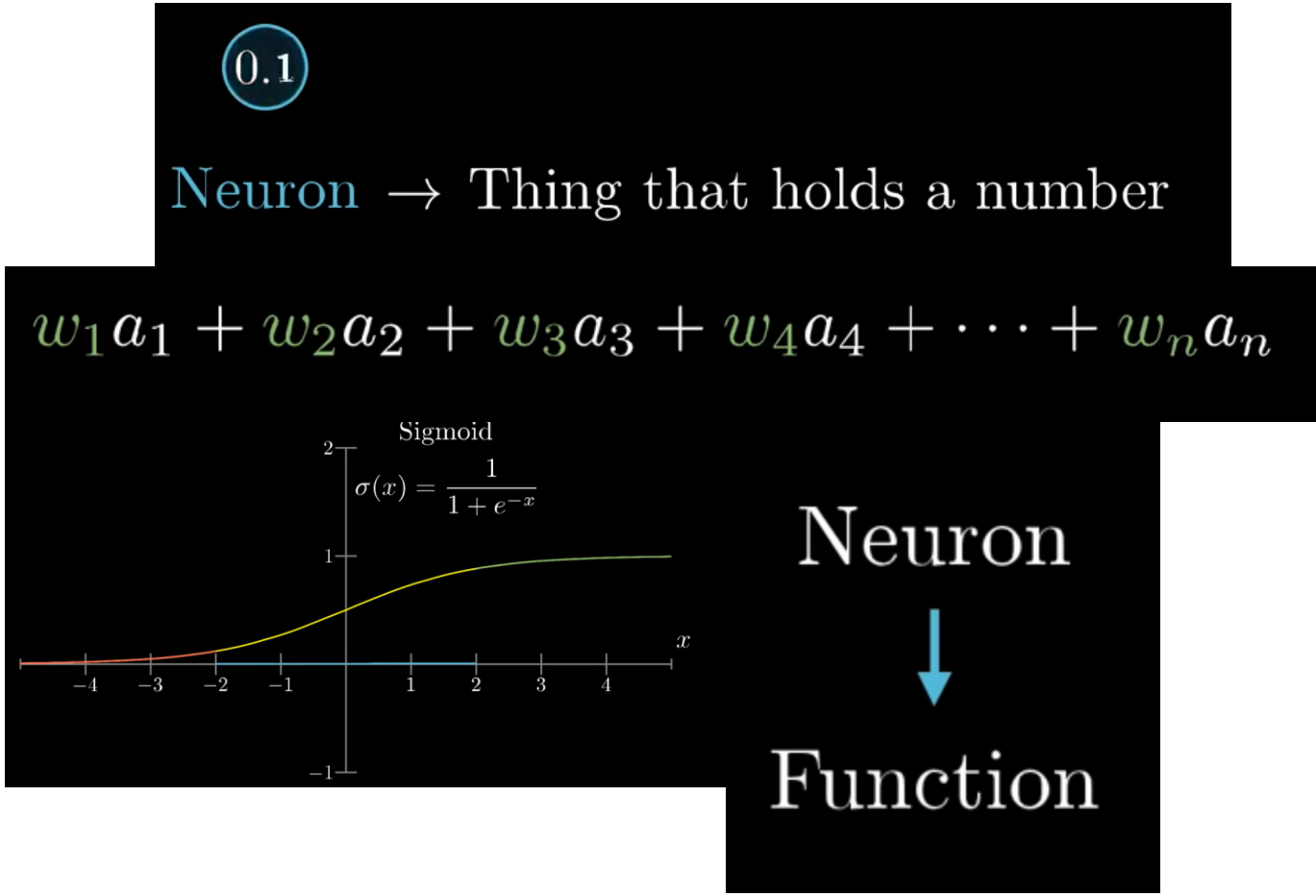
Masintõlke tööpõhimõtted

European Commission

(Tehis?)Neurovõrk







Kuidas masin keelt tõlgib?

1. Eeltöötlus (sisendi mõistmine)

- Segmenteerimine ja tokeniseerimine
- Sildistamine
- Alamsõnade segmenteerimine

2. Põhiline tõlketehnika
- Reeglipõhine masintõlge
- Statistiline masintõlge
- Neuromasintõlge
- LLM’id
- Hübriidsüsteemid

3. Konteksti ja ebaselguse käsitlemine (tähelepanumehhanism)

4. Järeltöötlus

- Grammatika ja süntaksiparandus
- Spetsiaalsete fraaside käsitlemine
- Detokeniseerimine

European Commission

Kuidas tõlkemootorit arendatakse?

Andmete kogumine

Andmete ettevalmistamine (puhastamine)

Mootori treenimine

Mootori optimeerimine

Mootori testimine ja hindamine

Mootori kasutuselevõtt

Masintõlke väljakutsed

European Commission

Masintõlke peamised väljakutsed

- Andmete piiratus
- Konteksti mõistmine
- Kultuurilised ja keelelised normid
- **Väikekeelte väljakutsed**

European Commission

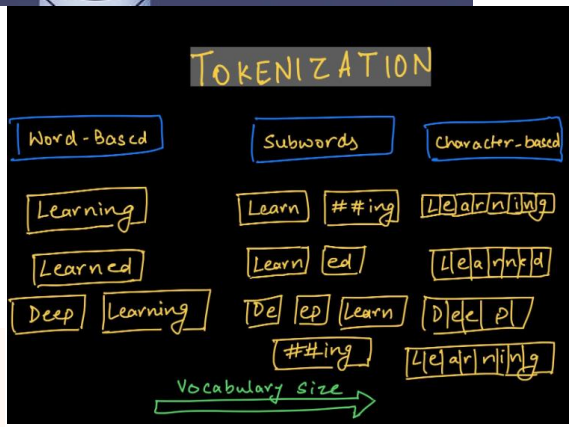
Sõnade välja mõtlemine

1. **Andmepiirangud**

2. Tundmatute sõnade käsitlemine

3. Liigne üldistamine

4. Tokeniseerimine



5. Lähtekeele ebamäärasus

6. Konteksti mittemõistmine

European Commission

Nimede "tõlkimine„ ja muutmine

1. Nimetuvastuse puudumine
2. Kultuuri- või keelenormid
3. Asendamine sarnase kõlaga nimedega

Maia → Maria

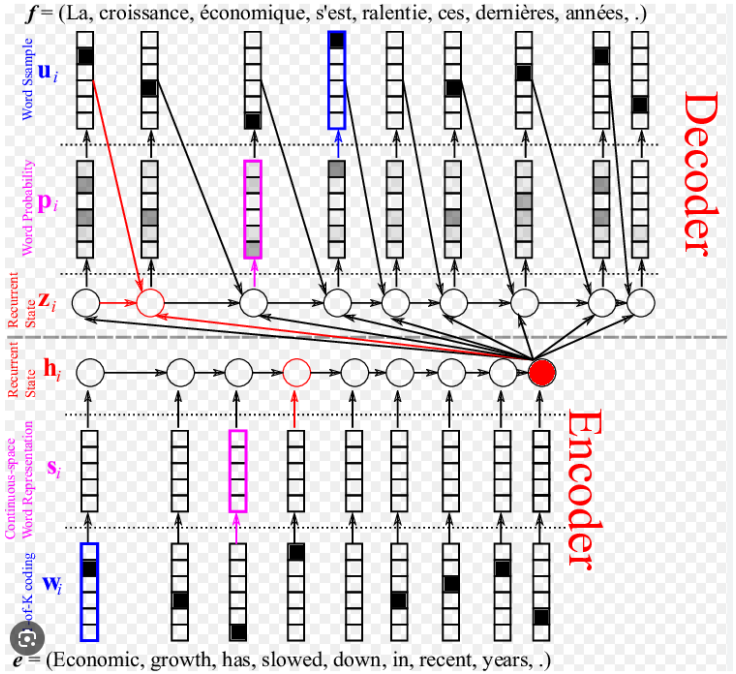
Nimi	masintõlge
Pille Rebane	
Mathias Juust	
Kadri Kull	

4. Nimede vale tõlgendamine nimisõnadena
5. Foneetiline tõlge
6. **Vead koolitusandmetes**

European Commission

Ühe sõna korrutamine

1. Mudeli ülesobitamine
2. Usalduslävendid
3. Keeruliste lausete käsitlemine
4. Konteksti mittemõistmine
5. Vigade taasloomine
6. **Andmete taskaalutus**



European Commission

Vahekokkuvõtteks

- Andmete piiratus – kättesaadavus, maht, kvaliteet.
- Konteksti mõistmine
- Kultuurilised ja keelelised normid
- **Väikekeelte väljakutsed**

European Commission

Eesti keele kui väikekeele väljakutsed 

European Commission

Eesti keele eripärad

Estonian open data

- Piiratud andmebaasid
- Keeruline morfoloogia
- Paindlik sõnajärjestus
- Ainulaadne sõnavara
- Vähene investeeringuhuvi
- Vähene kasutamine veebis

European Commission

Keeruline morfoloogia

- Käänetesüsteem ja mitmetähenduslikkus
- Konsonantmuutused ja tüve vaheldumine
- Keeruline mitmuse moodustamine
- Aglutinatsioon ja pikad sõnad
- Sõnajärje paindlikkus
- Morfeemide tõlkimise mitmetähenduslikkus
- Asesõnade käänded ja isikulised pöörded
- Piiratud andmed ja väikekeele staatus

European Commission

Paindlik sõnajärjestus

- Sõnade funktsiooni määramine
- Tõlkimisel puudub kindel sõnade järjekord
- Konteksti ja rõhuasetuse mõistmine
- Morfeemide ja käändevormide tugi

Mina andsin õpetajale raamatu.

Mina andsin raamatu õpetajale.

Raamatu andsin mina õpetajale.

Õpetajale andsin mina raamatu.

Õpetajale andsin raamatu mina.

Ainulaadne sõnavara

- Eesti keelele eripärased sõnad
- Liitsõnad ja tuletised
- Unikaalsed grammatilised konstruktsioonid
- Haruldased ja vähe levinud sõnad
- Neologismid ja arhailised sõnad

Masintõlke areng Euroopa Komisjonis

European Commission

eTranslation

European Commission

Kes saavad eTranslationit kasutada?

- Euroopa Liidu institutsioonide tõlgid ja muud töötajad

EL institutsioonid

- Avalik sektor ja riigiasutused
- Haridusasutused
- Väiksed ja keskmised ettevõtted
- Mittetulundusühingud
- Projektid, mida rahastatakse Digital Europe Programme poolt
- EPSO kandidaadid

EL (+Norra/Island/Liechtenstein/Ukraina)

European Commission

Kuidas eTranslationile ligi pääseda?

Kasutajad väljaspool EL instititutsioone:

https://webgate.ec.europa.eu/etranslation/public/welcome.html

Welcome to the registration page for eTranslation, the European Commission's safe and secure machine translation system. It also gives you access to the expanding range of AI tools developed under the Digital Europe programme. These include eSummary, Anonymization and more.

To use eTranslation and the other AI language tools, you need two things:

- an EU Login ID;
- an eTranslation account.

1. To register for an EU Login, click here . Once you have your EU Login credentials, return to this page to complete the next steps.
2. Enter the email address associated with your EU Login account:

3. eTranslation is intended for European public administrations, local and regional authorities, small and medium-sized enterprises, EU Freelance Translators, universities, non-governmental organizations and Digital Europe Programme projects. EPSO candidates are also eligible during the recruitment process.

Please indicate the type of user you are: ...

I certify that I belong to the category indicated above and therefore fall within eTranslation's remit.

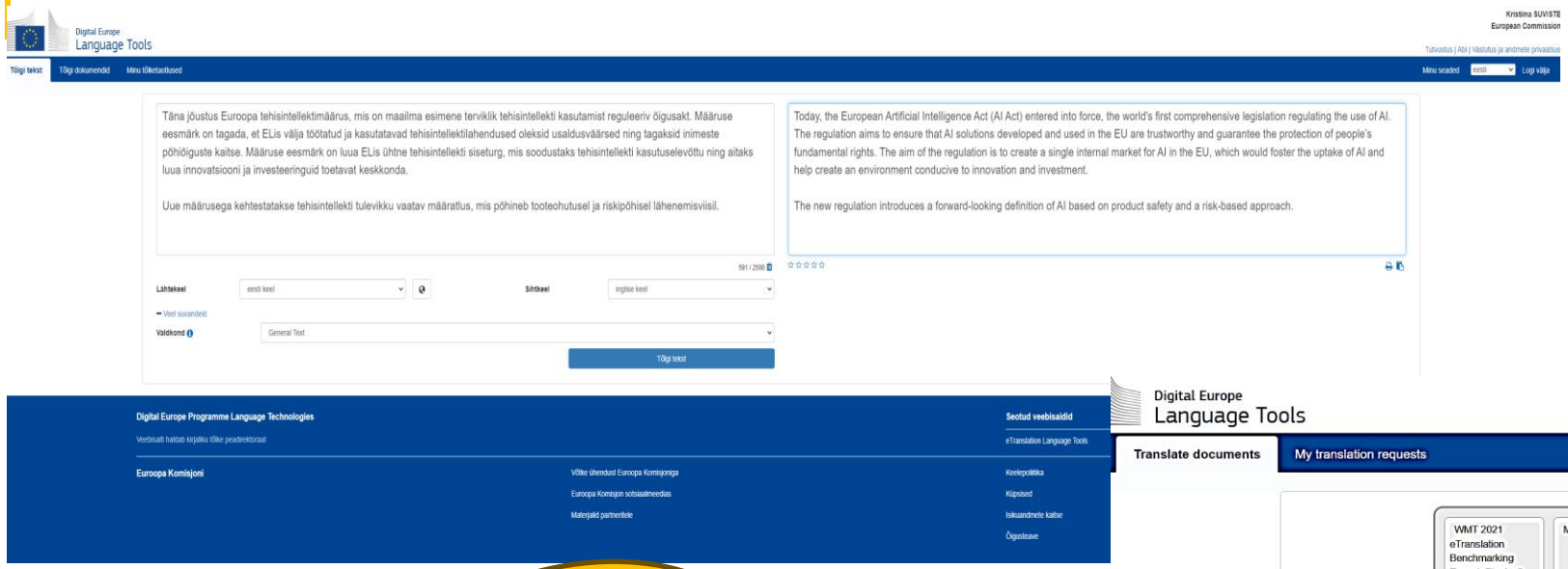
+ Privacy Statement

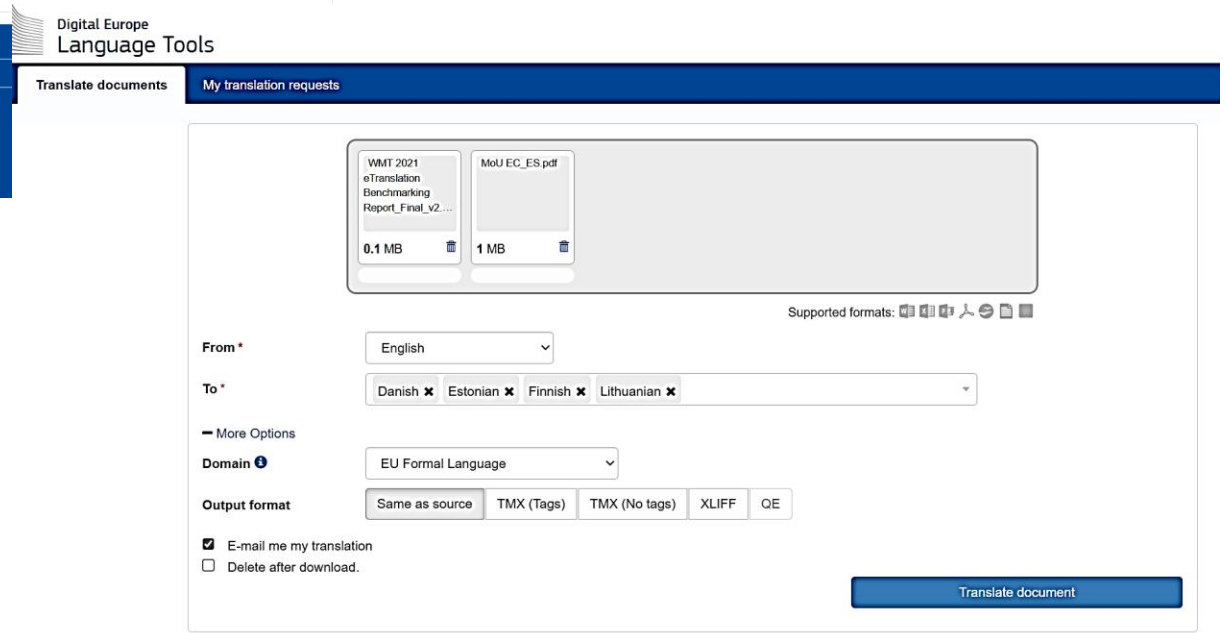
Complete your registration and start translating!

If you have any questions, please contact DGT-AI-Language-Services-Advisory.

European Commission

eTranslationi veebiliides



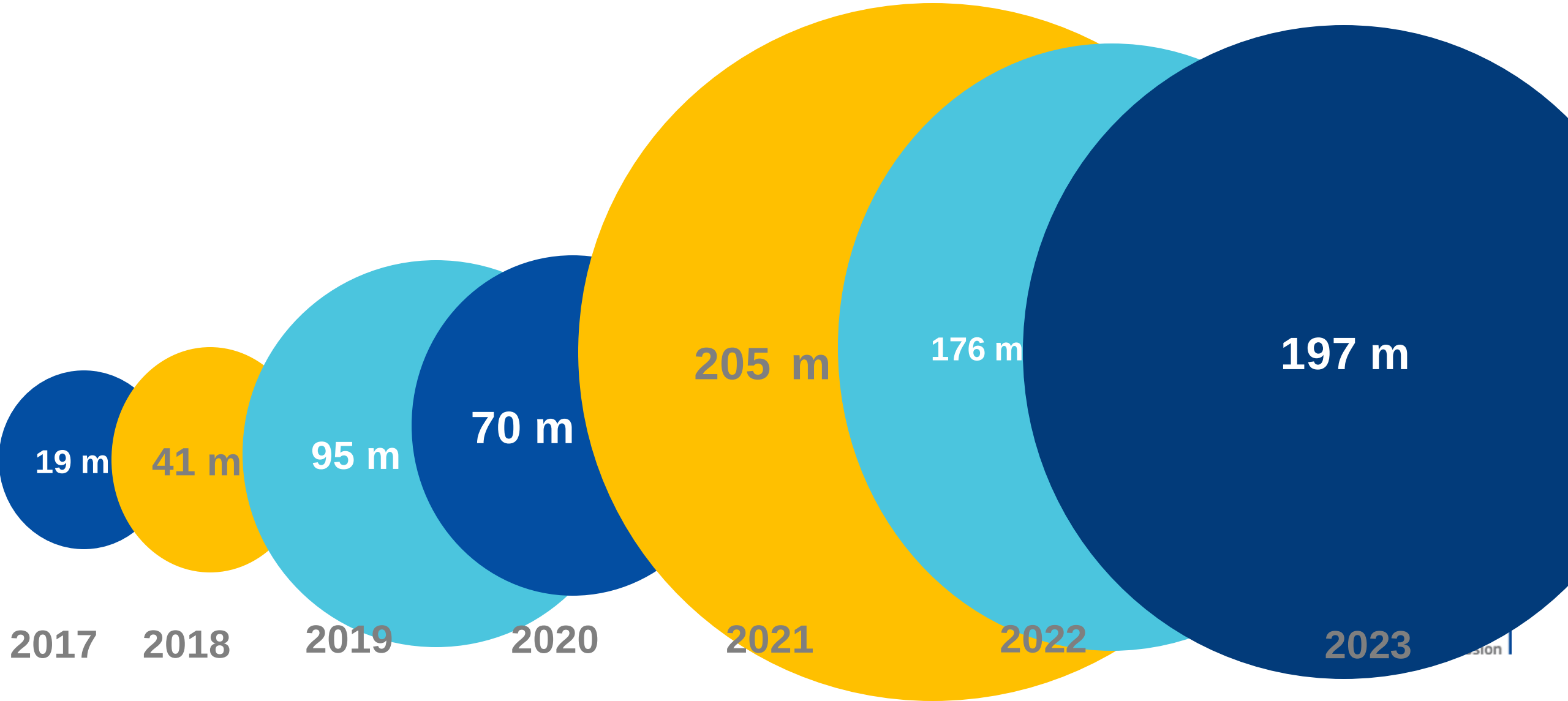


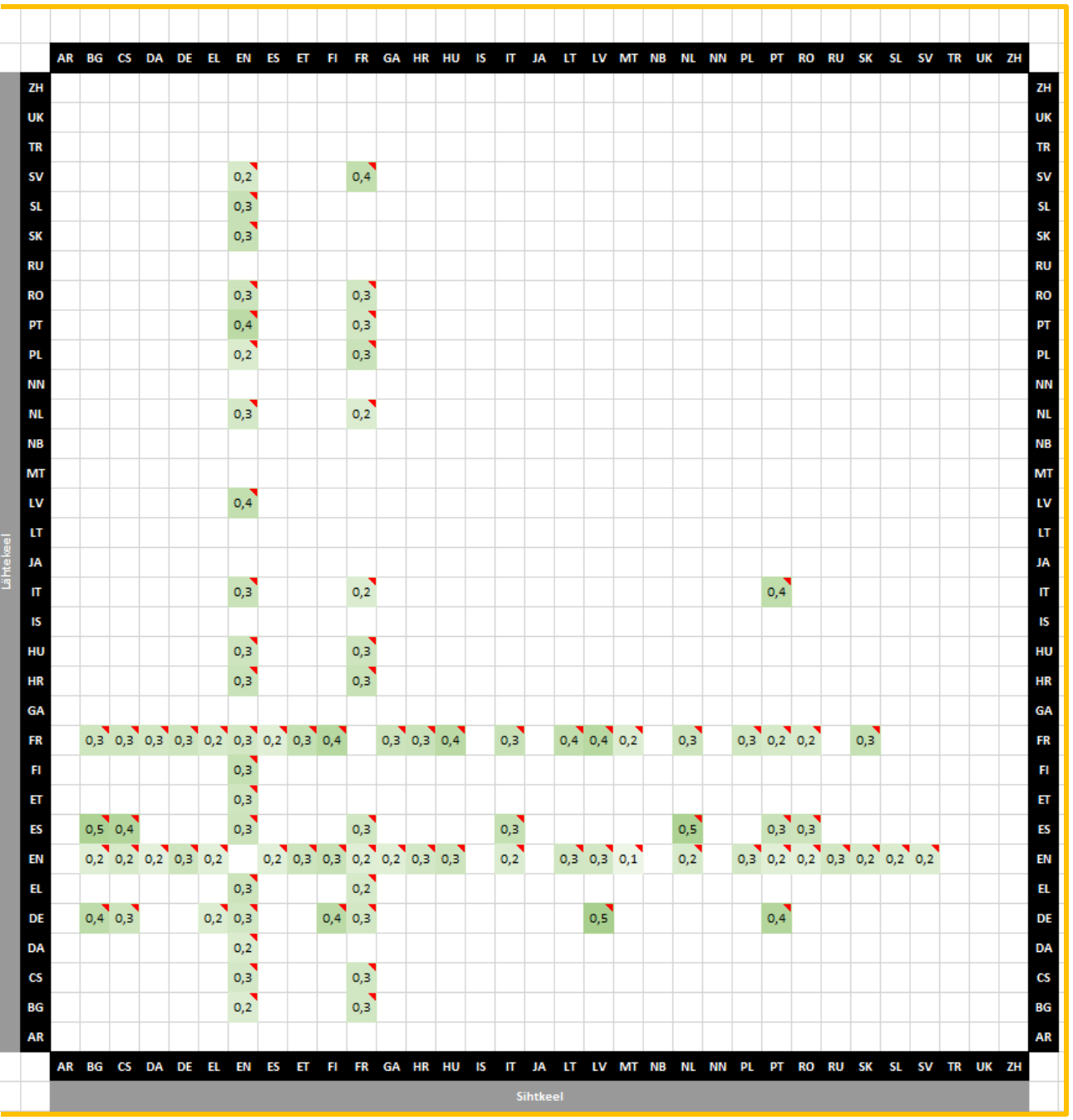
32 keelt saadaval

https://webgate.ec.europa.eu/etranslation

European Commission

eTranslationi kasutamine 2017-2023
(kõik kasutajad)





eTranslationi kvaliteet 2023

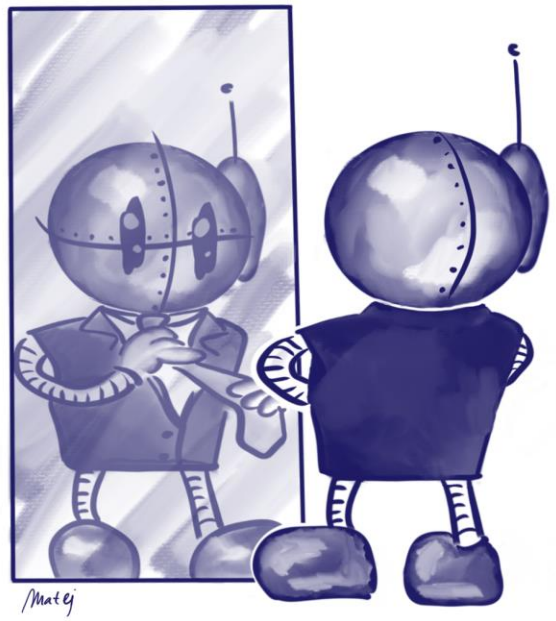
- Sõltub keelepaaridest, tekstitüüpidest
- TER *skoor* inglise keelest 0,2-0,4
- EN-ET 0,29; FR-ET: 0,34
- ET-EN 0,28

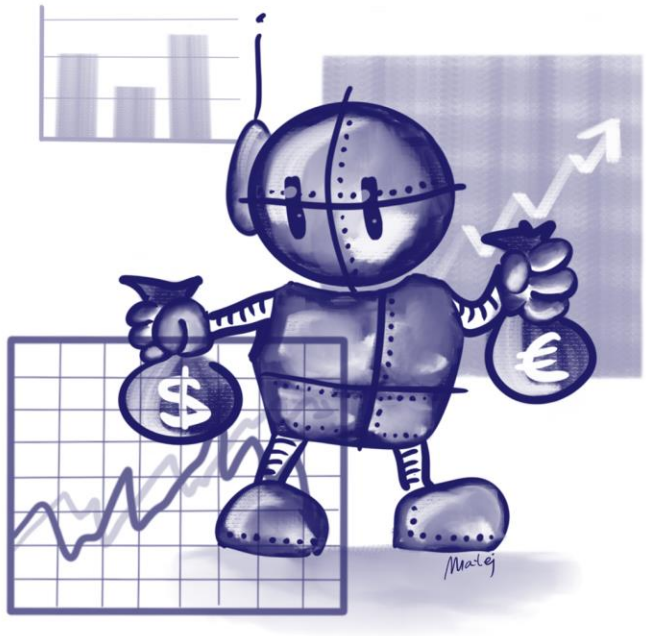
European Commission

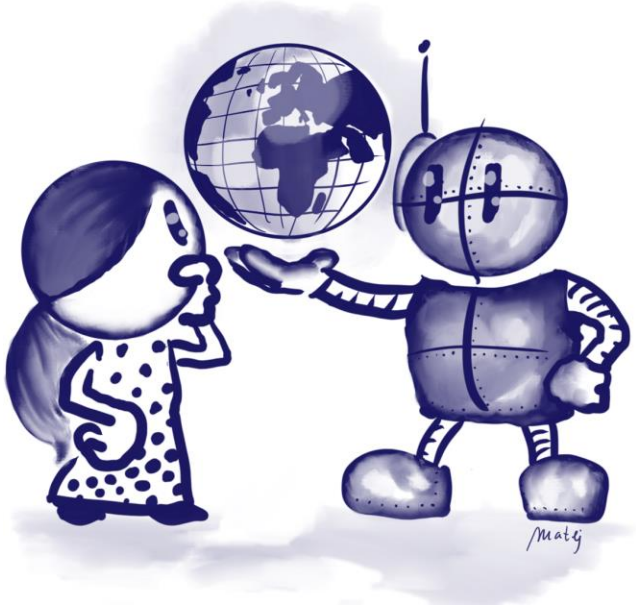
Lahendused masintõlke edasiseks arenguks

European Commission

Valdkonnapõhine tõlge







ope
ge Tools

ments | My translation requests

Click here to upload

Supported formats:

From *

To *

More Options

Domain

EU Formal Language

Output format

(No tags) | XLIFF | QE

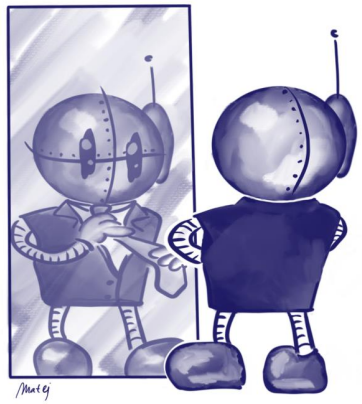
E-mail me my trans

Delete after downl

Translate document

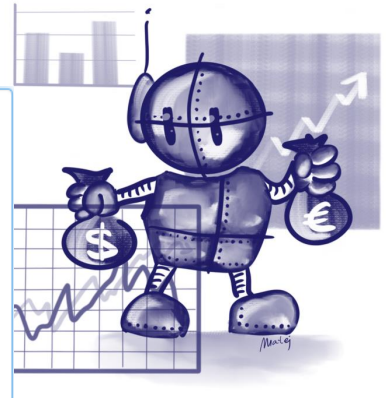
EU Formal Language
General Text
Court of Justice Case Law
Finance
IP Case Law
Public Health

https://webgate.ec.europa.eu/etranslation

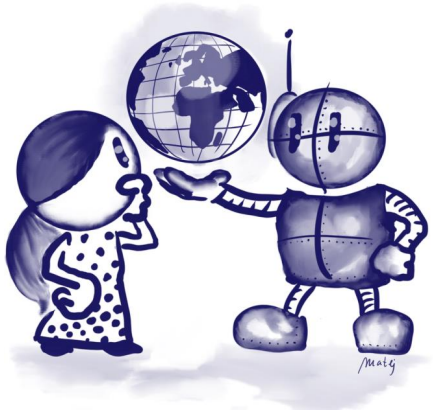


EU Formal

Masintõlkesüsteeme arendatakse suurtes kogustes kakskeelsete või mitmekeelsete tekstiandmete abil. Neid andmeid kasutatakse algoritmide treenimiseks, mis õpivad teksti tõlkimiseks ühest keelest teise. Tänapäevastes süsteemides, näiteks neurovõrkudel põhinevates süsteemides, kasutatakse süvaõppe meetodit, mille puhul õpetatakse mudeleid suurte andmestikega, et õppida keelemustreid ja -struktuure. Protsess hõlmab selliseid etappe nagu andmete kogumine, mudelikoolitus, testimine ja viimistlemine, et parandada tõlgete täpsust ja täpsust.

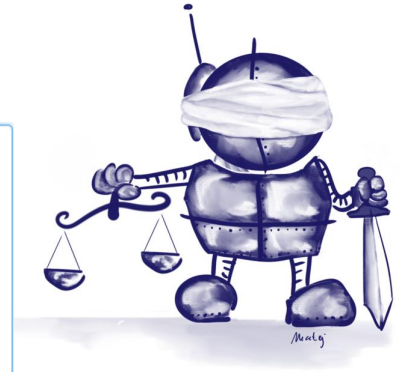


vs.



General

Masintõlkesüsteemid töötatakse välja suure hulga kaks- või mitmekeelsete tekstiandmete põhjal. Neid andmeid kasutatakse algoritmide treenimiseks, mis õpivad teksti ühest keelest teise tõlkima. Kaasaegsed süsteemid, nagu need, mis põhinevad närvivõrkudel, kasutavad meetodit, mida nimetatakse süvaõppeks, kus mudeleid koolitatakse tohutute andmekogumite abil, et õppida keele mustreid ja struktuure. Protsess hõlmab selliseid etappe nagu andmete kogumine, mudelikoolitus, testimine ja peenhäälestamine, et parandada tõlgete täpsust ja sujuvust.



European Commission

Image from adobe stock

EU Formal - **EURAMIS**

- **EL institutsioonidevaheline koostöö**
- **Hõlmab kahte kümnendit tõlkide tööst**
- Umbes **2 miljardit** lauset
- **400 000** uut lauset lisatakse iga päev

European Commission

Glossary: Environmental Policy & Climate Change

English Term	Estonian Translation
Sustainable Development	Jätkusuutlik areng
Greenhouse Gas Emissions	Kasvuhoonegaaside heitkogused
Renewable Energy	Taastuvenergia
Carbon Neutrality	Süsinikuneutraalsus
Climate Change Mitigation	Kliimamuutuste leevendamine
Circular Economy	Ringi- või ringmajandus
Biodiversity Conservation	Bioloogilise mitmekesisuse kaitse
Environmental Impact Assessment	Keskkonnamõjude hindamine
Paris Agreement	Pariisi leping
Carbon Trading	Süsinikukaubandus
Sustainable Agriculture	Jätkusuutlik põllumajandus
Energy Efficiency	Energiatõhusus
Deforestation	Metsade raadamine
Emission Reduction Targets	Heitkoguste vähendamise eesmärgid
Renewable Energy Sources	Taastuvenergiaallikad
Climate Adaptation Strategies	Kliimamuutustega kohanemise strateegiad
Waste Management	Jäätmekäitlus
Environmental Legislation	Keskkonnaalane seadusandlus
Public Awareness Campaign	Avalikkuse teadlikkuse tõstmise kampaania
Ecosystem Services	Ökosüsteemi teenused
Environmental Sustainability	Keskkonna jätkusuutlikkus
Renewable Energy Transition	Üleminek taastuvenergiale
Pollution Control Measures	Saaste kontrollimeetmed
Ecological Footprint	Ökoloogiline jalajälg
Sustainable Urban Development	Jätkusuutlik linnade areng

Terminibaasi lisamise võimalus eTranslationile

Oktoober 2024?

2 keelt

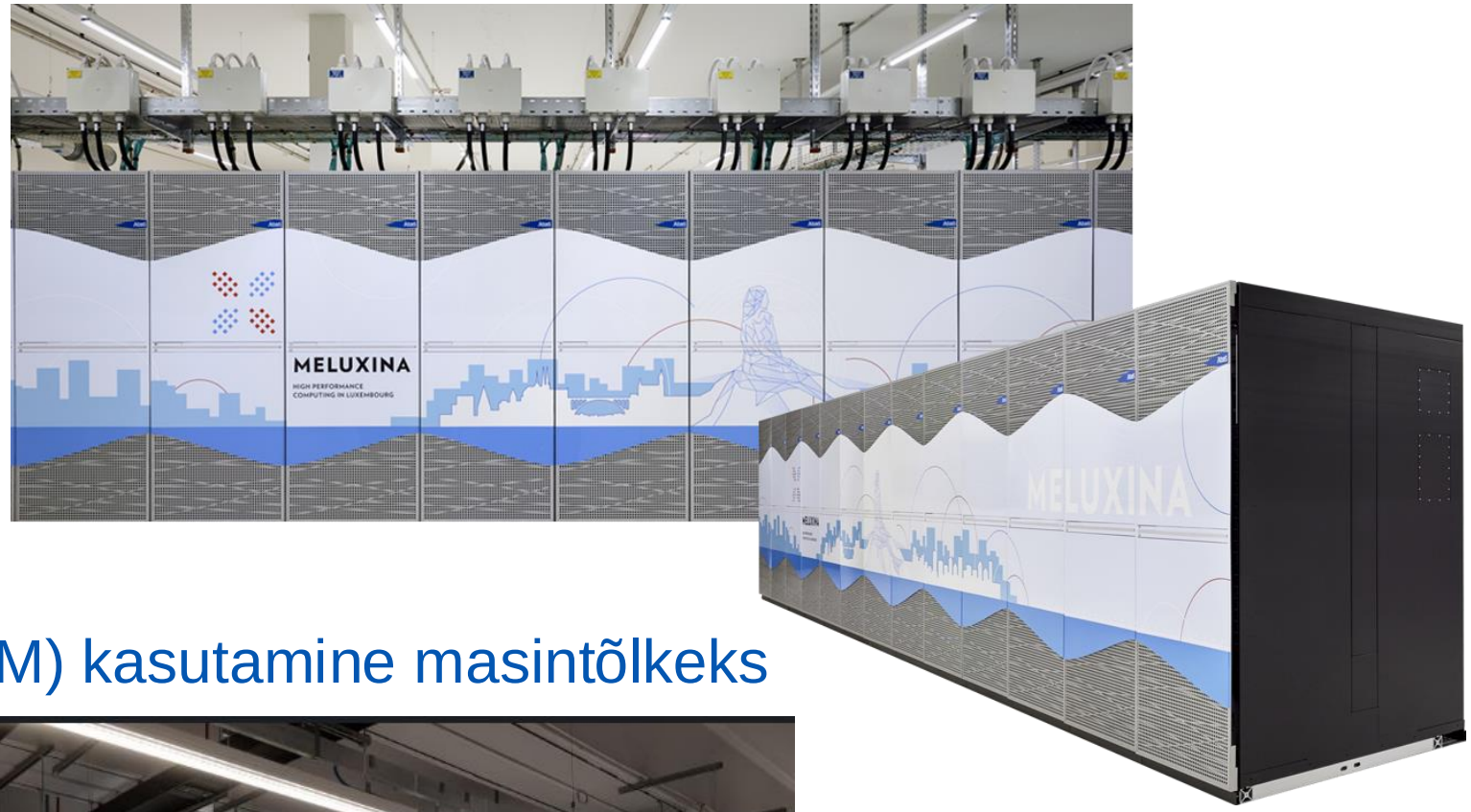
Lühikesed terminid

Max 250 sissekannet

European Commission

Suured keelemudelid

- Superarvutid



- Suurte keelemudelite (LLM) kasutamine masintõlkeks

European Commission

„EU LLM“ projekt – ANDMED

EURAMIS

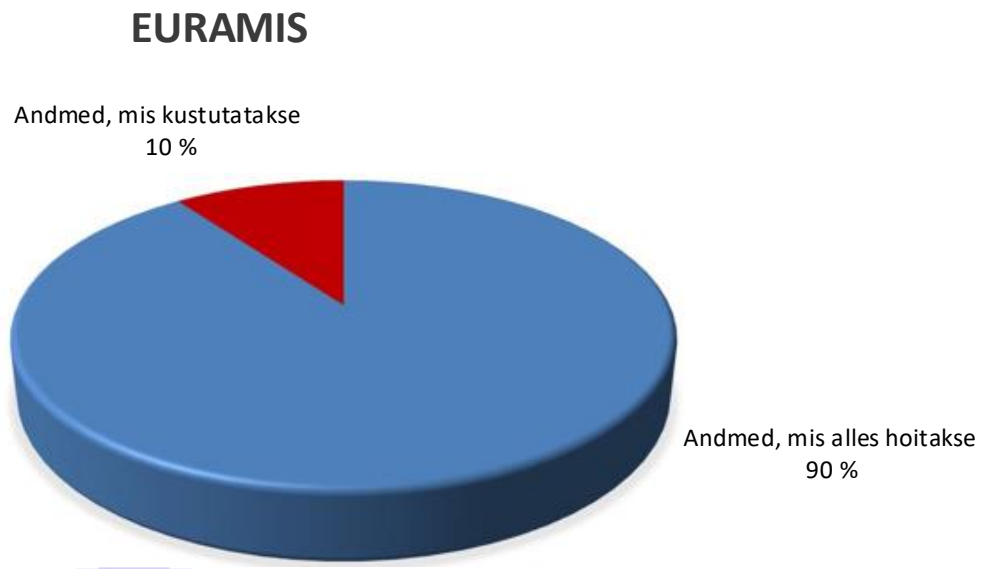
EL institutsioonide andmebaas

24

keelt

100B

tokenit



European Commission

„EU LLM“ projekt – treenitav mudel

MISTRAL AI_

8X7

Mistral AI | Frontier AI in your hands

3 x suurem kui Llama2-13B

EuroHPC kaudu kasutada:
100B tokenit - 50 000 GPU tundi

EURAMIS

European Commission

DGT ‘EU LLM’ projekt – võimalikud tulevikustsenaariumid

- Mitmekeelne LLM (kõik 24 EL keelt)
- Avatud VKE, haridusasutustele ja avalikule sektorile liikmesriikides
- Aluseks DGT TI-põhistele teenustele
- Võimsamad LLMid tulevikus
- Nullist treenimine?
- Panustada Euroopa tulevikukindlana hoidmisse

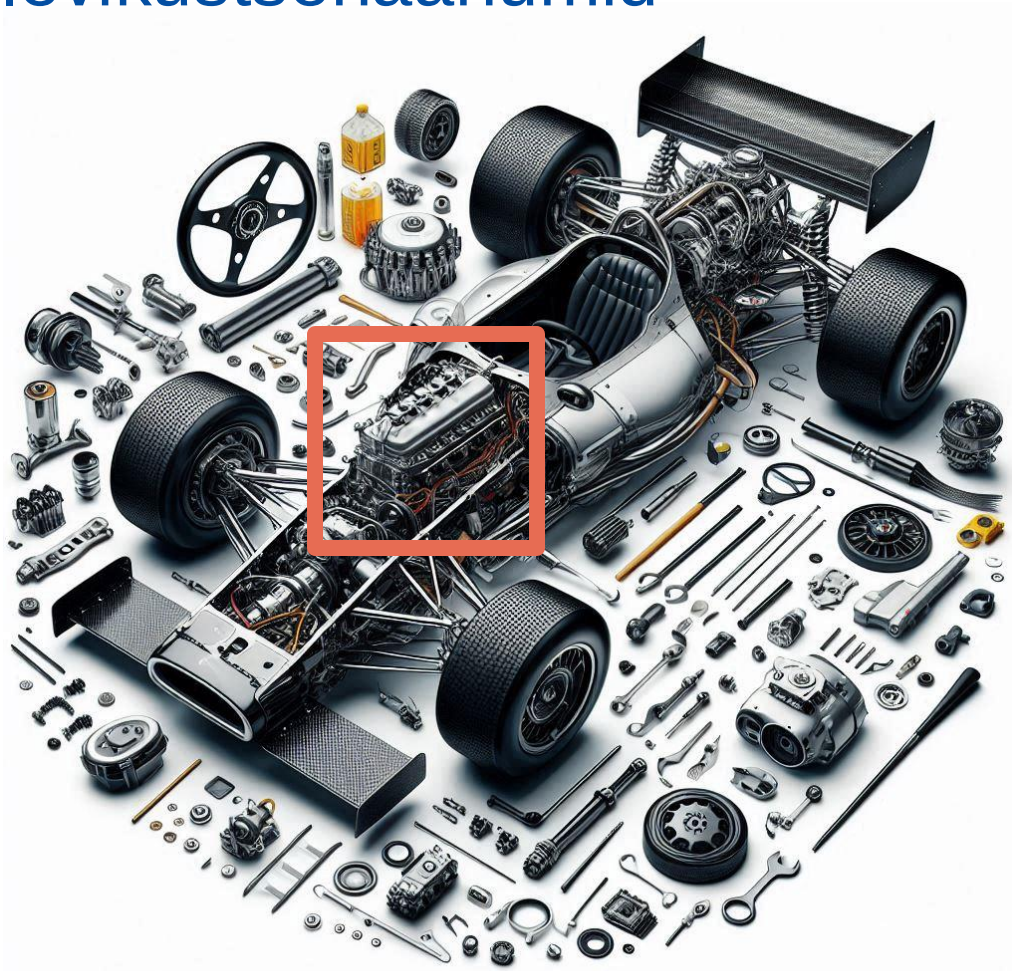


Image created by Copilot.

European Commission

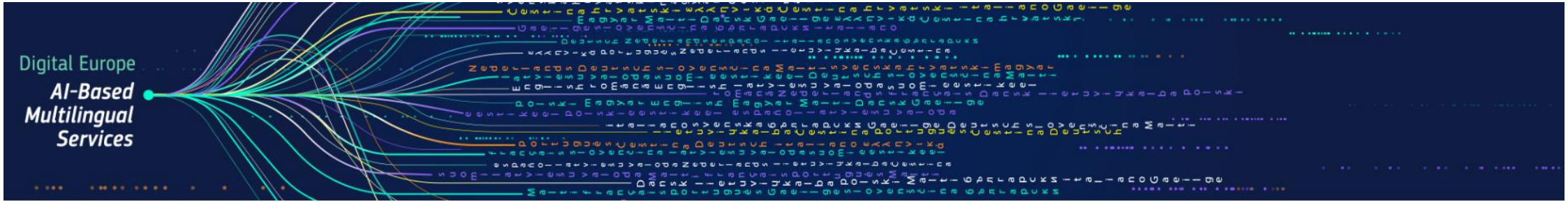
Küsimused?

DGT-AI-Language-Services-Advisory@ec.europa.eu

Tehisintellektil põhinevate keeleteenuste platvorm

European Commission

Veel teenuseid, mida pakume…



AI-Based Multilingual Services

These services are made available in the EU under the Digital Europe Programme for use by EU institutions, public administrations, academia, SMEs, NGOs, Digital Europe Programme projects and EPSO candidates.

They offer both web pages and APIs for machine-to-machine access.

eTranslation

Neural machine translation built on the EU's history of professional translation.

eBriefing

Generate reports from sets of documents in official or general styles.

eReply

Have AI help you prepare replies to correspondence, queries, and other requests.

eSummary

Quickly find out the main content of long documents.

Multilingual Post

Short translations in multiple languages in one shot for X.

Speech-to-Text

Upload your media and get full transcriptions or subtitles back.

Natural Language Processing Tools (NLP)

Anonymisation, Classification and Named-Entity Recognition.

Access and registration

Access to these tools requires registration.

EU staff is pre-registered

- Who can access these services?
- Registration page
- API registration page

Contact point:
DGT-AI-Language-Services-Advisory@ec.europa.eu

Additional useful resources

Developer's corner

Guidelines and procedures on how to integrate eTranslation in your web pages (only in English)

Web translation tool

Free plugin for your website

Language Data Space

European platform for exchanging languages resources

IATE

Interactive Terminology for Europe

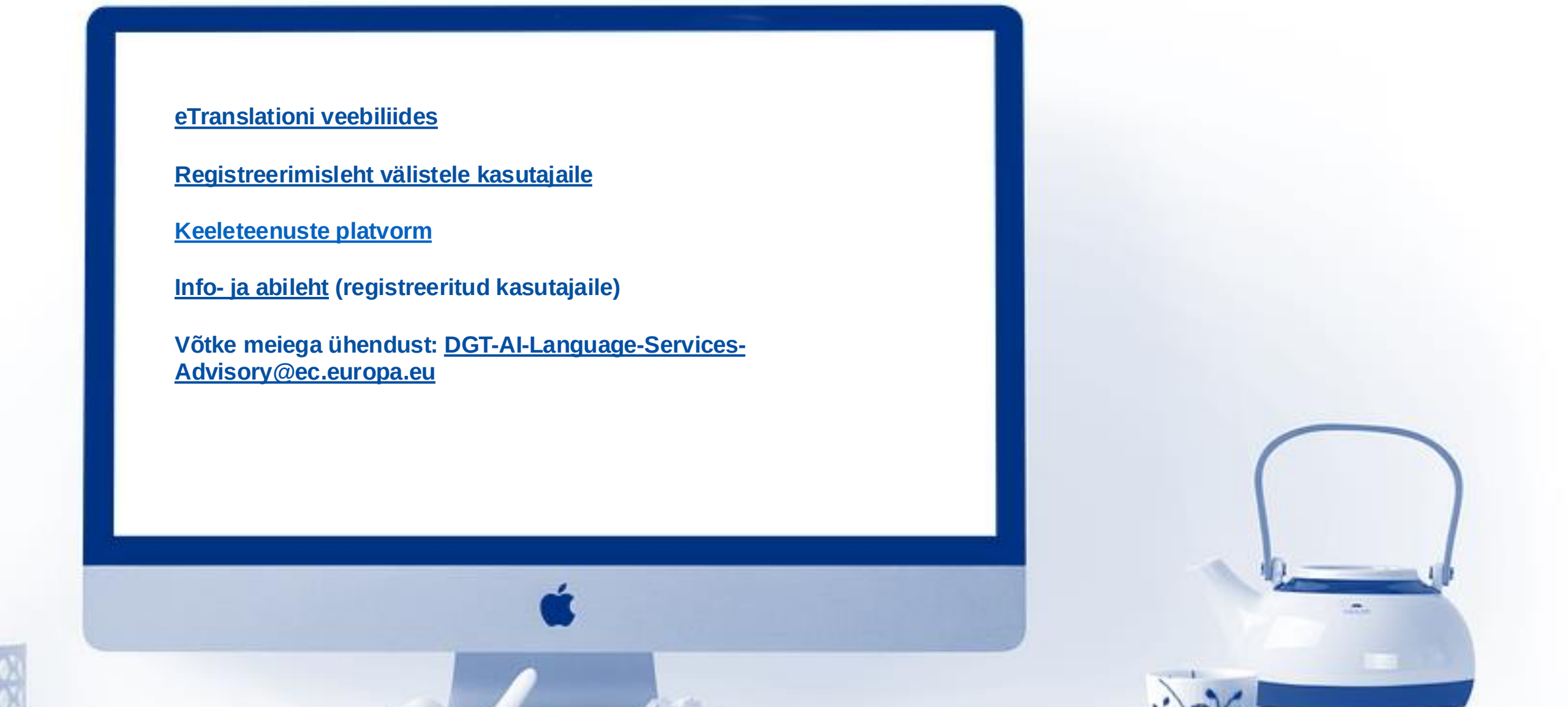
The Digital Europe Programme

- What is the **Digital Europe Programme** ?
- Questions and answers
- Follow the Commission's work on tech and digital **@DigitalEU**

[AI-Based Multilingual Services]

European Commission

Kasulikud lingid



eTranslationi veebiliides

Registreerimisleht välistele kasutajaile

Keeleteenuste platvorm

Info- ja abileht (registreeritud kasutajaile)

Võtke meiega ühendust: DGT-AI-Language-Services-Advisory@ec.europa.eu