



EUROPEAN LANGUAGE GRID

European Language Grid

An Introduction and Overview

Katrin Marheinecke (DFKI) – ELG Project Manager (katrin.marheinecke@dfki.de)

15-12-2020 – 4th Regional ELG Workshop – Finland

<http://www.european-language-grid.eu>

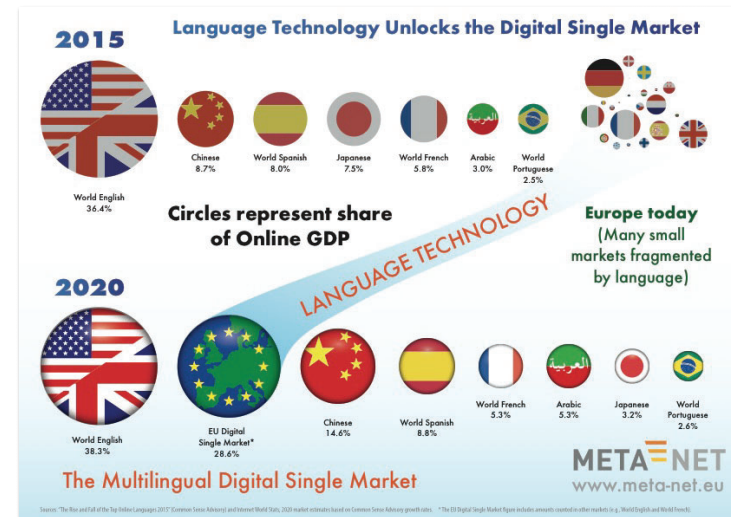
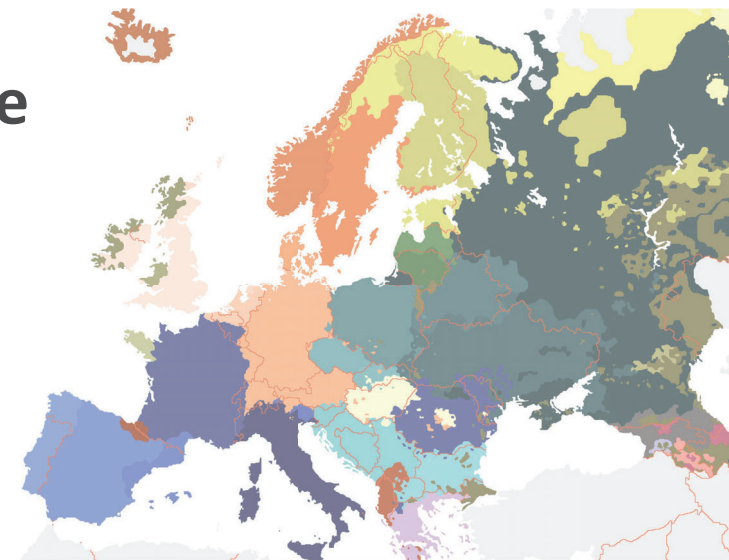
Coordinator: Dr. Georg Rehm (DFKI)

4th Regional ELG Workshop: Finland

14:00	Welcome and introduction	Krister Lindén (University of Helsinki)
14:05	ELG: history and overview	Katrin Marheinecke (DFKI)
14:30	ELG: online demo	Nils Feldhus (DFKI)
14:50	Finnish Pilot Projects funded in ELG	Filip Ginter (University of Turku) Sebastian Andersson (Lingsoft) Jörg Tiedemann (University of Helsinki)
15:20	Finnish Language Platform Initiative	Marko Turpeinen (1001 Lakes)
15:40	Summary and discussion	
16:00	End of the workshop	

Point of Departure: Multilingualism in Europe

- Multilingualism is at the heart of the European idea
- 24 official EU languages – they all have the same status
- Dozens of co-official, regional and minority languages as well as languages of immigrants and trade partners
- Many economic, social and technical challenges
 - The Digital Single Market needs to be multilingual
 - Cross-border, cross-lingual, cross-cultural communication
 - Fragmentation of the LT market and landscape



Language Technology is already everywhere

- Language Technology makes use of theoretical results of language-oriented research to create applications and technological solutions. Fields and areas involved:
 - Artificial Intelligence + Computer Science
 - Computational Linguistics
 - Natural Language Processing (NLP)
 - Natural Language Understanding (NLU)
 - Psychology, Psycholinguistics
 - Cognitive Science
- **Language Technology = *Language-centric AI***

- Spell and grammar checker in MS Word
- Web search (Google, Bing, Yandex etc.)
- Social Media Analytics, Media Monitoring
- Voice control for phones or computers
- Voice control in cars
- Machine translation
- Conversational agents and chatbots (Echo, Home, Siri, Cortana etc.)
- Speech synthesis in games
- Computer-Assisted Language Learning (CALL)
- Optical Character Recognition (OCR)

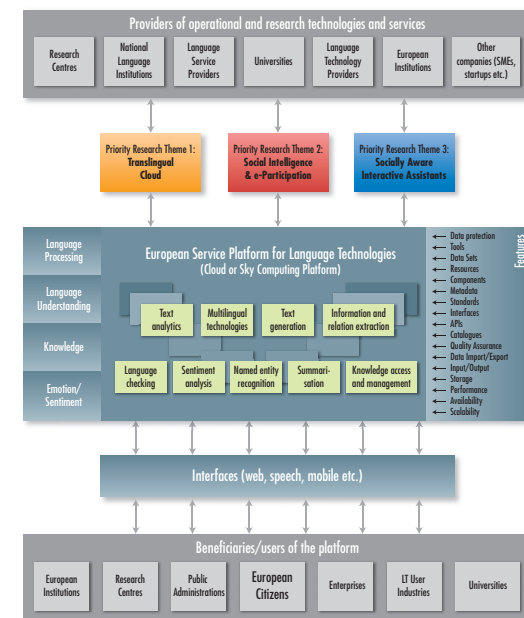
Selected applications that include Language Technology

Motivation and Context

- The European LT landscape and market are very fragmented.
- The European LT community has been demanding a **European Language Technology platform** for several years.
- **META³NET** vision: **European Service Platform for Language Technologies**.
- First mention in the **Strategic Research Agenda for Multilingual Europe 2020** (published in early 2013).



META³ FORUM 2013



The concept was later refined in the three LT SRIAs (2015, 2016, 2017).



META³ FORUM 2015



META³ FORUM 2016



META³ FORUM 2017

European LT Market – CEF Study

Approach and Observations

- Background: SMART 2016/0103 contract: contribute to CEF AT as “multilingualism enabler” for CEF DSIs.
- European LT vendors grouped per type of tech: Translation, Speech, Search, NLU, Analytics
- **EU market approx. 1B€ in 2020 – disrupted by dominant global players**
- SMEs: 70% of EU LT vendors up to 50 employees
- Revenue per company is growing
- Market is highly fragmented: hundreds of SMEs, many address very specific niches, sectors and languages

Recommendations

- Europe is strong in R&I, but not successful to scale innovations and capture the market
- Europe needs European alternatives to fill the gaps and deficiencies and to avoid reliance on monopolies
- Multilingual DSM should be developed on its own infrastructure
- Public procurement can be the major driver for European LT industry to avoid dependence on monopolies
- Plans needed to avoid brain drain
- *A platform is needed to connect demand and supply as well as industry and research*

➤ **CEF study estimates the European LT Market to reach 1B€ in 2020.**



The Global Natural Language Processing Market size is expected to reach \$29.5 billion by 2025, rising at a market growth of 20.5% CAGR during the forecast period

Globe Newswire | FOLLOW+
January 14, 2020 4:26am | Comments

➤ **Truly incredible market forecast!**
(Source: Global Newswire)



European Language Grid

ELG Project



Kick-off meeting, 22/23 January 2019

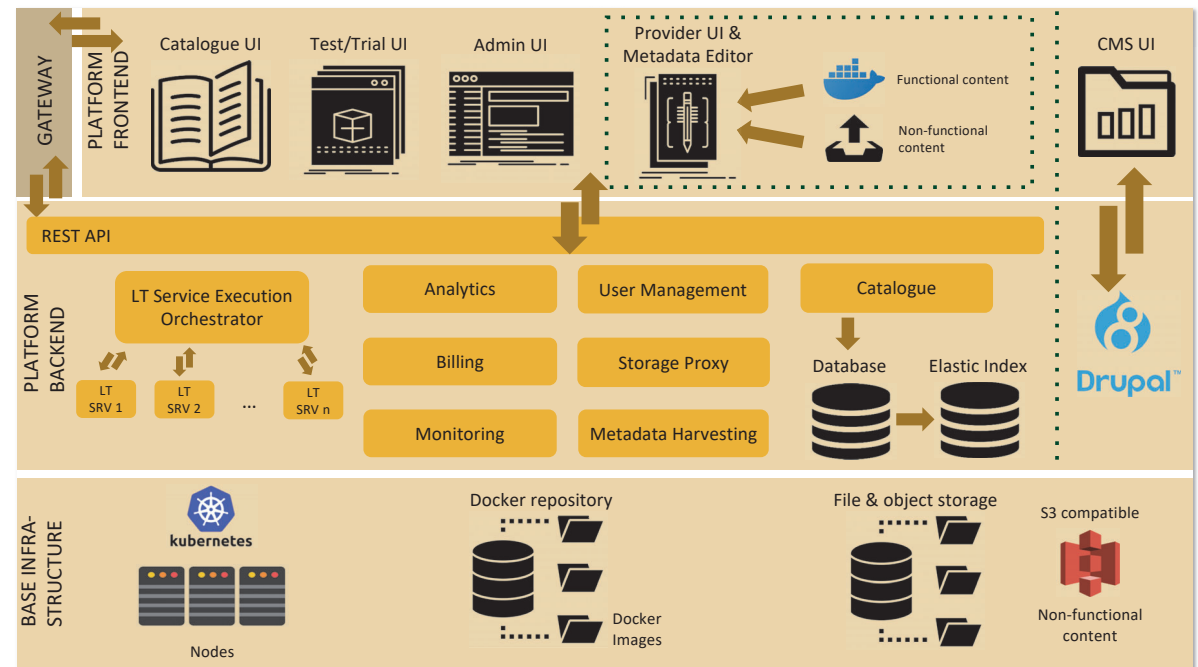
Objectives (Selection)

1. Establish the ELG as the primary Language Technology platform and market place in Europe to tackle the fragmentation of the European LT landscape.
2. ELG as a platform for commercial and non-commercial, industry-related LTs, both functional and non-functional.
3. Enable the European LT community to upload services and data sets into the ELG, to deploy them and to connect with, and make use of those resources made available by others.
4. Enable businesses to grow and benefit from scaling up.
5. Unleash enormous potential for innovation.

European Language Grid – Current State of Play (December 2020)

- User registration, authentication, authorisation
- User categories, respective rights and policies
- LT metadata upload and editing facilities
- Metadata conversion and harvesting
- LT service registration, integration
- LT service try out and execution
- LT data browsing, searching, upload, download
- Online documentation
- Current state of play in the European Language Grid towards ELG Release 2 (early 2021):
 - 168 functional services and tools
 - 1998 corpora
 - 729 lexical/conceptual resources
 - 7 language descriptions

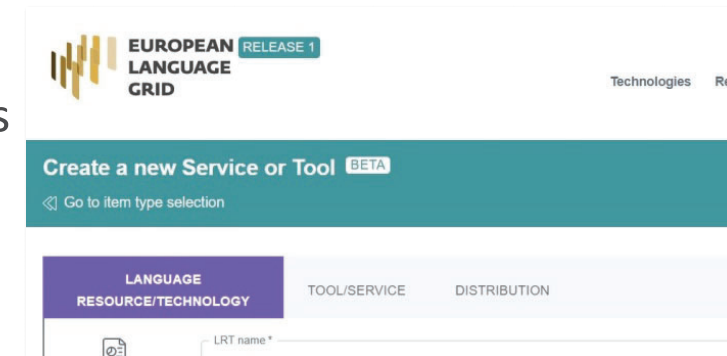
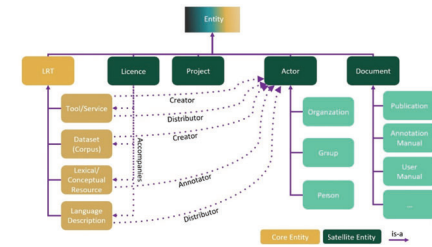
- *Facilities for consumers, esp. of data and services*
- *Facilities for providers, esp. of data and services*



Users can connect to the ELG cloud platform via ELG APIs, remote APIs, ELG GUI, Python client library, download of containers or source code.

European Language Grid – What can you do with it?

- **Data consumers** can search and browse the ELG catalogue
 - for different types of data, language processing services, projects and organisations
 - download data (depending on access conditions)
- **Service consumers** can try out and test language processing services
 - call a service from the command line, integrate it into workflow
 - view code samples
 - current API support: MT, IE, ASR, TTS, text classification
- **Data and service consumers** can use a Python-based API for accessing the ELG catalogue, searching and directly fetching datasets to feed them into, e.g., their model training pipeline
- **Data providers** can upload resources and register metadata descriptions using a semantic schema
 - a dashboard provides an overview of own resources including their status (*draft* → *published*)
- **Language Technology providers** can provide their service/tool as a Docker container
 - Different options: Docker image directly in ELG; Docker image in ELG talks to remote service



LT Services: Current state of play

- **ELG Release 1** (April 2020) finalised APIs for major classes of services (ASR, IE, MT, TTS)
 - Concentrated on Czech, English, French, German, Greek, Latvian, Spanish (native ELG languages)
 - 150 distinct services for Text Analytics: 52 English, 28 German, 21 French, 14 Greek
 - 24 MT services
 - 9 ASR services
 - 2 TTS (Latvian, Lithuanian)
- **ELG Release 2** (early 2021) will add support for other EU and related languages, at least:
 - 8 further ASR, 200-250 further Text Analysis, 23 further MT, 9 further TTS services
 - We also expect the first services and datasets from the ELG pilot projects
- **ELG Release 3** (early 2022)
 - Services for an even wider range of non-EU languages, current projection is for at least
 - 15 further ASR, 160 further Text Analysis, 9 further MT services
 - Additional service types: image OCR, terminology extraction from corpora, etc.

	Language Group			Totals
	A	B	C	
Basic linguistic processing	146	30	63	239
Entity recognition and linking	102	7	11	120
Classification (sentiment, opinion, topic, language etc.)	40	4	18	62
Machine Translation	105	4	14	123
Other text processing (including parsing, summarisation etc.)	35	7	13	55
Speech tools (ASR, gender detection, synthesis etc.)	26	5	14	45
Totals	454	57	133	644

Services and tools to be ingested by the ELG partners (according to the proposal)

ELG: Data Sets and Language Resources

	Corpora	Lexical and Conceptual Resources	Models and Computational Grammars	Total
ELRA	635	545	–	1180
ELRC-SHARE	844	43	–	887
META-SHARE	52	12	7	71
LINDAT/CLARIAH-CZ	243	66	–	309
ELRA-SHARE	46	25	–	71
Zenodo	36	37	–	73
Total	1857	727	7	2591

	Open Access	Language Group A			Language Group B			Language Group C			Totals
		Corpora	Lexicons	Models	Corpora	Lexicons	Models	Corpora	Lexicons	Models	
META-SHARE	yes	617	447	16	55	54	0	84	51	1	1325
	no	582	550	1	44	65	0	198	94	0	1534
ELRC-SHARE	yes	317	114	0	3	1	0	0	0	0	435
	no	74	16	0	2	1	0	0	0	0	93
ELDA	no	563	1012	0	35	18	0	250	54	0	1932
ELG	mixed	74	108	43	0	0	12	4	1	21	263
Totals		2227	2247	60	139	139	12	536	200	22	5582

Datasets and resources to be provided by the ELG partners (according to the proposal)



European Language Grid



**EUROPEAN
LANGUAGE
GRID**

RELEASE 1

Languages

Official EU Languages

- + English (1823)
- + Spanish (514)
- + German (395)
- + French (378)
- + Czech (269)

Show more

Other EU/European languages

- + Russian (72)
- + Turkish (58)
- + Norwegian Bokmål (51)
- + Catalan (50)
- + Basque (48)

Show more



**EUROPEAN
LANGUAGE
GRID**

RELEASE 1

Language resources & technologies

- + Corpus (1998)
- + Lexical/Conceptual resource (729)
- + Tool/Service (168)
- + Language description (7)

A – Official EU languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish

B – EU candidate languages and Free Trade Partners: Albanian, Basque, Catalan, Galician, Icelandic, Norwegian, Scottish Gaelic, Welsh, Serbian, Turkish, Ukrainian

C – Languages spoken by EU immigrants; languages of political and trade partners: Afrikaans, Arabic, Berber, Cebuano, Chinese, Hebrew, Hindi/Urdu, Indonesian, Japanese, Korean, Kurdish, Latin, Malay, Pashto, Persian (Farsi), Russian, Tamil, Vietnamese

Stakeholders and Users

Companies that

- ... *develop* Language Technologies
- ... *integrate* Language Technologies
- ... *purchase* Language Technologies

Universities and research centres that

- ... *develop* Language Technologies
- ... *use* Language Technologies

Public administrations that *purchase* or *use* Language Technologies

Other organisations (e.g., NGOs) that *purchase* or *use* Language Technologies

Funding agencies that support the development of Language Technologies



META-FORUM 2019 (8/9 October) – Brussels, Belgium

META-FORUM Conference Series

META-FORUM 2020 – December 01-03, *virtual conference*

Piloting the European Language Grid

META-FORUM 2019 – October 08/09, Brussels, Belgium

Introducing the European Language Grid

META-FORUM 2017 – November 13/14, Brussels, Belgium

Towards a Human Language Project

META-FORUM 2016 – July 04/05, Lisbon, Portugal

Beyond Multilingual Europe

META-FORUM 2015 – April 27, Riga, Latvia

Technologies for the Multilingual Digital Single Market

META-FORUM 2013 – September 19/20, Berlin, Germany

Connecting Europe for New Horizons

META-FORUM 2012 – June 20/21, Brussels, Belgium

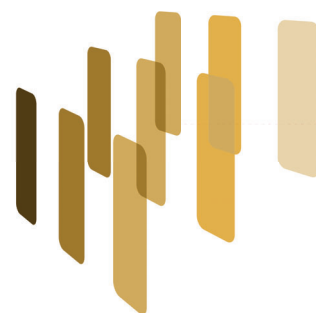
A Strategy for Multilingual Europe

META-FORUM 2011 – June 27/28, Budapest, Hungary

Solutions for Multilingual Europe

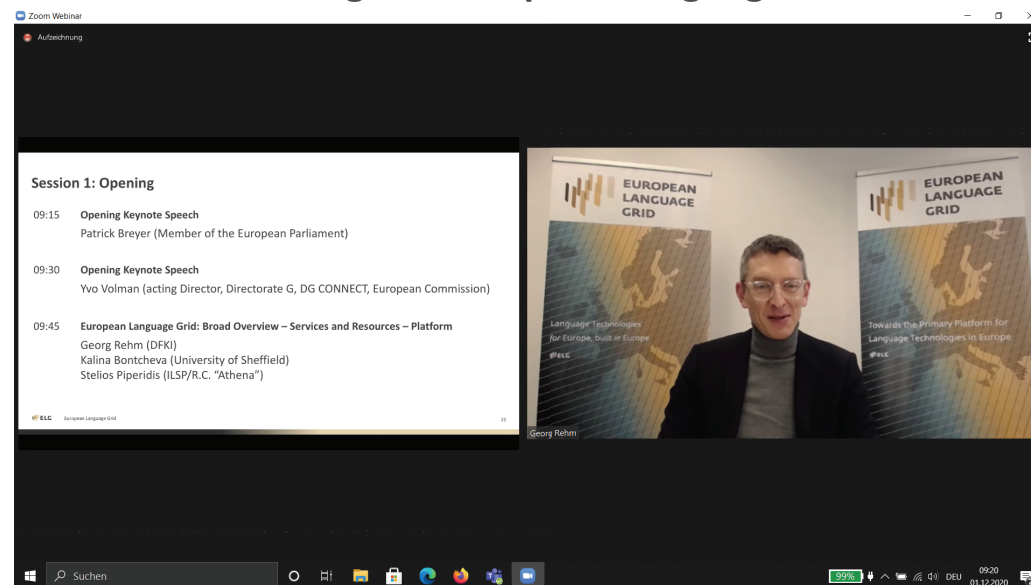
META-FORUM 2010 – November 17/18, Brussels, Belgium

Challenges for Multilingual Europe



EUROPEAN LANGUAGE GRID

META-FORUM 2020: Second Annual ELG Conference Piloting the European Language Grid





European Language Grid

META-FORUM 2019



Community: NCCs and LTC

32 National Competence Centres (NCCs)

- Strong international network of national networks to broaden ELG's reach, identify content for the ELG and interest companies in using the ELG.
- Main goal: *support the mission of the ELG project.*

European LT Council (LTC)

- A pan-European body, in which strategic LT-related matters can be discussed and coordinated.
- Main goal: *represent and support European LT community.*



Austria



DAGMAR GROMANN

Centre for Translation Studies, University of Vienna

Belgium



WALTER DAELEMANS

Computational Linguistics and Psycholinguistics, University of Antwerp

Bulgaria



SVETLA KOEVA

Institute for Bulgarian Language, Bulgarian Academy of Sciences

Croatia



MARKO TADIĆ

Department of Linguistics, University of Zagreb

Cyprus



DORA LOIZIDOU

University of Cyprus

Czech Republic



JAN HAJIČ






Institute of Formal and Applied Linguistics, Charles University Prague

Denmark



BOLETTE SANDFORD PEDERSEN

Centre for Language Technology, Department of Nordic Studies and Linguistics, University of Copenhagen

1. **Austria:** Dagmar Gromann, Zentrum für Translationswissenschaft, Universität Wien
2. **Belgium:** Walter Daelemans, Comp. Ling. and Psycholing. Res. Centre (CLiPS), University of Antwerp
3. **Bulgaria:** Svetla Koeva, Bulgarian Academy of Sciences
4. **Croatia:** Marko Tadic, Inst. of Linguistics, Faculty of Hum. and Social Science, University of Zagreb
5. **Cyprus:** Dora Loizidou, University of Cyprus
6.  **Czech Republic:** Jan Hajic, Inst. of Formal and Applied Linguistics, Charles University in Prague
7. **Denmark:** Bolette Sandford Pedersen, Centre for Lang. Technology, Department of Nordic Research, University of Copenhagen
8. **Estonia:** Kadri Vare, Department of Language, Estonian Ministry of Education and Research
9. **Finland:** Krister Lindén, Department of Modern Languages, University of Helsinki 
10. **France:** François Yvon, CNRS-LIMSI
11.  **Germany:** Georg Rehm, Speech and Language Technology Lab, DFKI
12.  **Greece:** Maria Gavrilidou, ILSP, R.C. "Athena"
13. **Hungary:** Tamás Várádi, Research Institute for Linguistics, Hungarian Academy of Sciences
14. **Iceland:** Eiríkur Rögnvaldsson, School of Humanities, University of Iceland
15. **Ireland:** Andy Way, ADAPT Centre and School of Computing, Dublin City University
16. **Italy:** Bernardo Magnini, Human Language Technology, Fondazione Bruno Kessler (FBK)
17. **Latvia:** Inguna Skadina, Institute of Mathematics and Computer Science, University of Latvia
18. **Lithuania:** Albina Aukšoriūtė, Institute of the Lithuanian Language
19. **Luxembourg:** Dimitra Anastasiou, Luxembourg Institute of Science and Technology
20. **Malta:** Michael Rosner, Department Intelligent Computer Systems, University of Malta
21. **Netherlands:** Jan Odiijk, Utrecht Institute of Linguistics, Universiteit Utrecht
22. **Norway:** Kristine Eide, The Language Council of Norway – Språkrådet
23. **Poland:** Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
24. **Portugal:** António Branco, Department of Informatics, University of Lisbon
25. **Romania:** Dan Tufis, Research Institute for Artificial Intelligence, Romanian Academy of Sciences
26. **Serbia:** Cvetana Krstev, Faculty of Mathematics, Belgrade University (UBG)
27. **Slovakia:** Radovan Garabik, Ludovít Stur Institute of Linguistics, Slovak Academy of Sciences
28. **Slovenia:** Simon Krek, Jozef Stefan Institute
29. **Spain:** Marta Villegas, Barcelona Supercomputing
30. **Sweden:** Jens Edlund, Speech, Music & Hearing/Språkbanken Tal, KTH Royal Institute of Technology
31. **Switzerland:** Hervé Bourlard, Idiap Research Institute
32.  **UK:** Kalina Bontcheva, Department of Computer Science, University of Sheffield



European Language Grid: Sustainable Operational Model & Legal Entity

- ELG is supposed to be a long-term, sustainable initiative – a legal entity is needed.
- The technical and operational requirements – high availability and performance, SLAs, billing, support etc. – create non-trivial costs: hosting; bandwidth; ELG team; legal; etc.
- We've identified several ways and approaches of covering the costs on a long-term basis.
- Establish consensus for a sustainable operational model.
- Options: a) for-profit or b) not-for-profit company, c) association, d) foundation
- First concept prepared (Business Model Canvas).
- Q4/2021: Establish legal entity (probably with a soft start)



Open Calls for Pilot Projects

Two open calls for pilot projects

- **Open Call #1:** 03/04 2020
- **Open Call #2:** 10/11 2020 ←

Pilot projects shall

- **Type A:** broaden ELG's portfolio or
- **Type B:** demonstrate usefulness of ELG

Up to €200,000 per project

Approx. €2,000,000 FSTP in total

Available in Open Call #1: 1.3M€

Available in Open Call #2: 585k€ ←

Project duration: 9-12 months

Eligibility: SMEs, research orgs.

Open Call #1 – Statistics !

110 project proposals evaluated

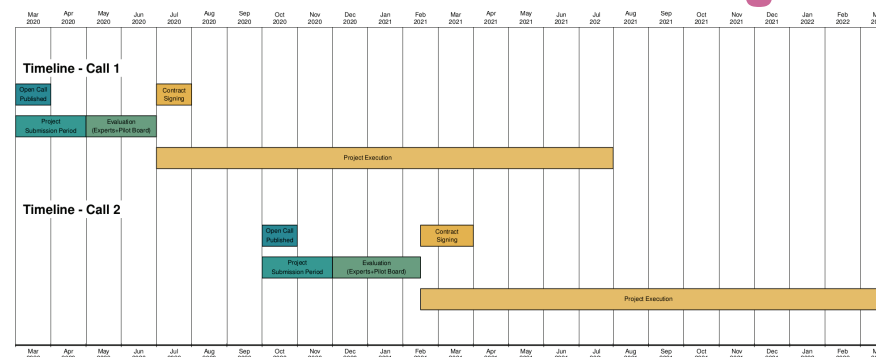
- **Type A:** 79 proposals
- **Type B:** 31 proposals

Total amount requested: 16.9M€

10 projects selected on 29 June 2020

Open Call #2 – Statistics !

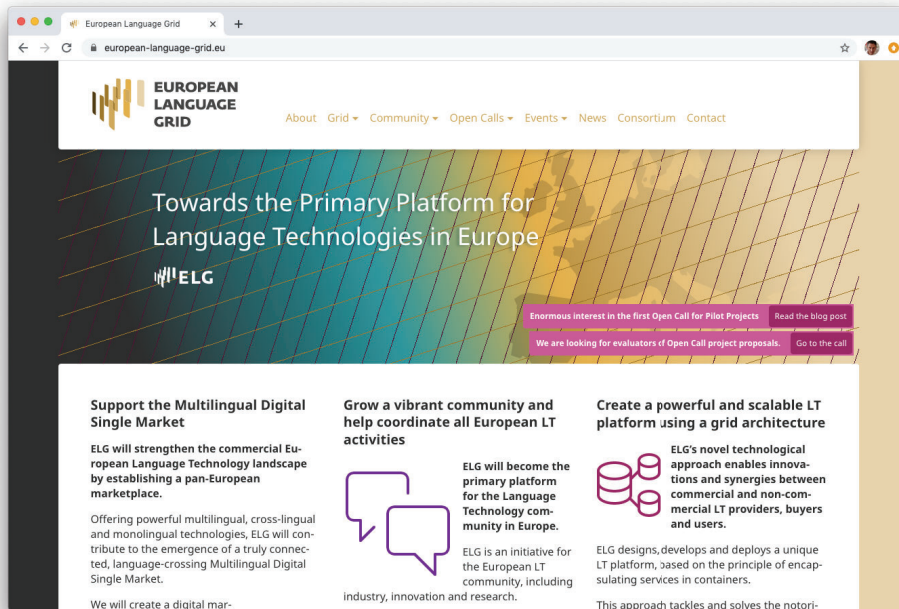
106 project proposals submitted



Open Call #1: Selected Pilot Projects

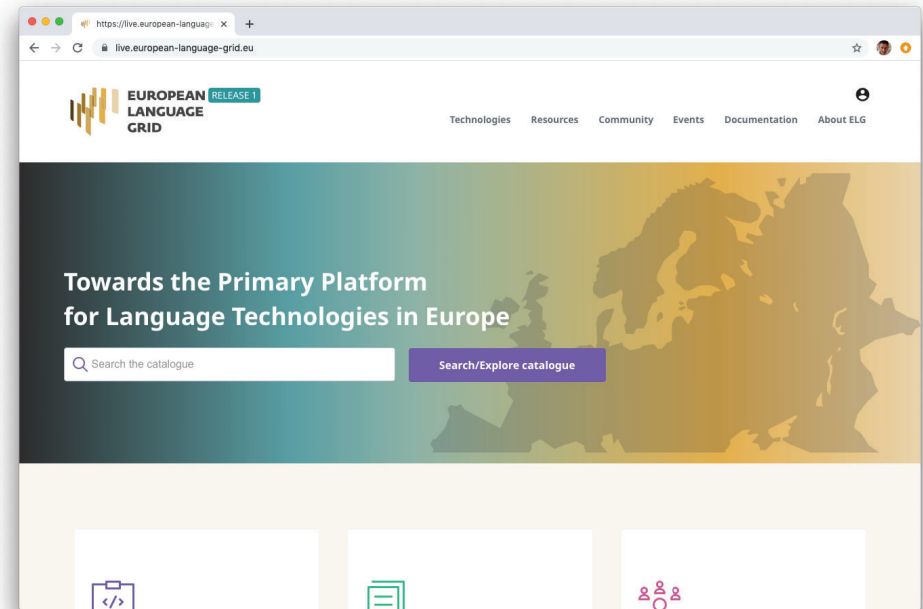
Fondazione Bruno Kessler	European Clinical Case Corpus	Italy	EUR 139,370
Lingsoft, Inc.	Lingsoft Solutions as Distributable Containers	Finland	EUR 140,625
Coreon GmbH	MKS as Linguistic Linked Open Data	Germany	EUR 167,375
Elhuyar Fundazioa	Basque-speaking smart speaker based on Mycroft AI	Spain	EUR 117,117
Universita' Degli Studi di Torino	Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools [...]	Italy	EUR 126,125
University of Helsinki	Open Translation Models, Tools and Services	Finland	EUR 154,636
Centre for Translation Studies, University of Vienna	Extracting Terminological Concept Systems from Natural Language Text	Austria	EUR 132,977
University of Turku, Turku NLP research group	Textual paraphrase dataset for deep language modelling	Finland	EUR 166,085
Weber Consulting KG	Virtual Personal Assistant Prototype	Austria	EUR 87,445
FZI Research Center for Information Technology	Streaming Language Processing in Manufacturing	Germany	EUR 132,160

European Language Grid



<https://www.european-language-grid.eu>

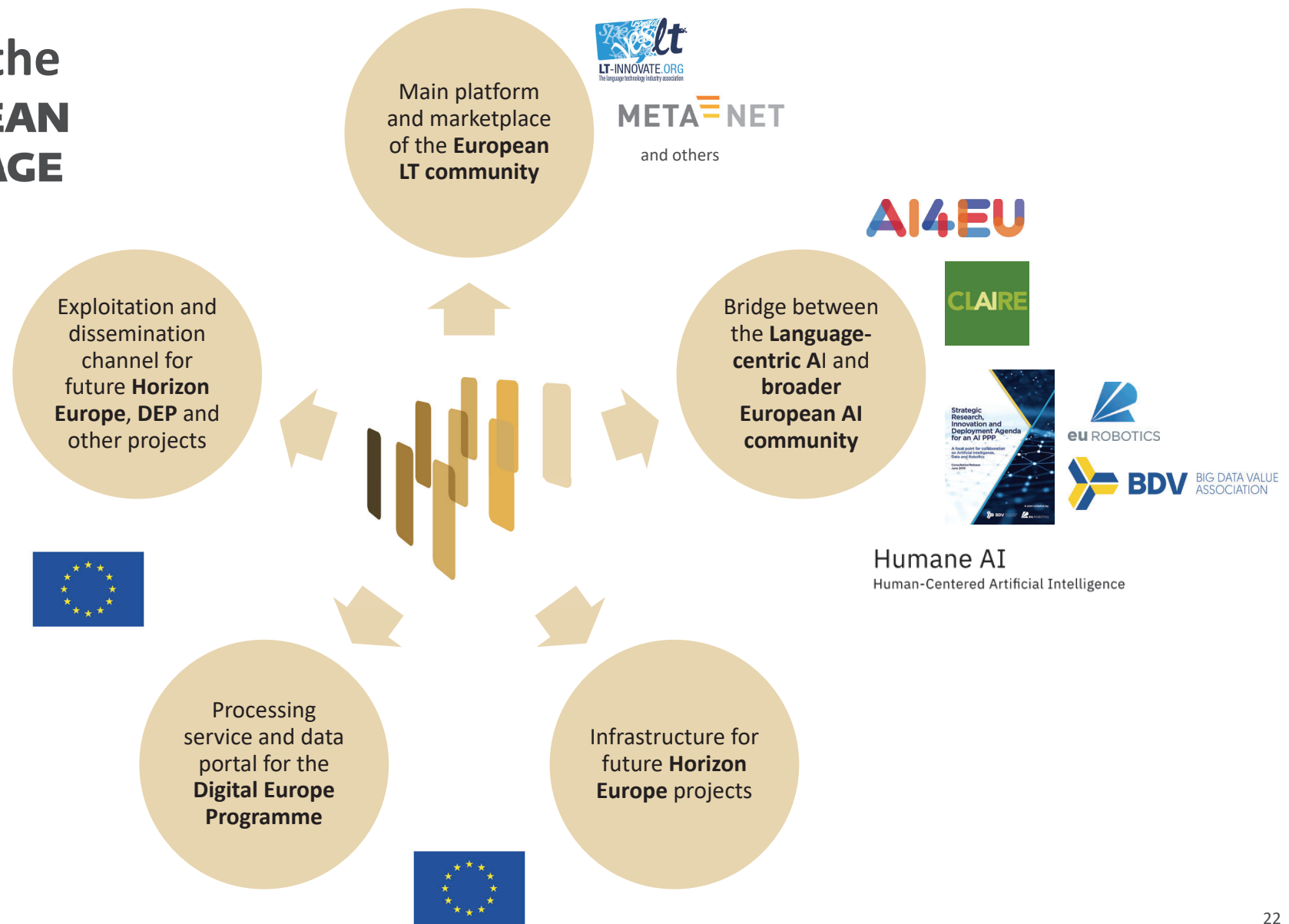
- ELG project website
- Typical EU project information plus a lot of additional content (events, open calls, community, NCCs etc.)
- The whole content of the project website will be integrated into the European Language Grid website (proper) in 2021



<https://live.european-language-grid.eu>

- The actual European Language Grid platform and website
- The next revision of the ELG project website (left) will include dynamic content that will be pulled from the ELG repository (right), e.g., languages, number of services, categories of services etc.

Future roles of the EUROPEAN LANGUAGE GRID



Summary and Next Steps

- Establish ELG as the primary platform and marketplace for Language Technology in Europe.
- An initiative *from* the European LT community *for* the European LT community.
- European LT landscape is **highly fragmented**: ELG aims to provide just the right **umbrella platform**.
- Global market size by 2025 is enormous: we want the European LT community to be a **key player**.
- We want to **increase the visibility and reach** of all members of the European LT landscape.
- ELG is a long-term initiative: we will establish a **legal entity** for sustainability.
- Contribute to the long-term goal of **Digital Language Equality** in Europe by giving all our languages one virtual home and umbrella platform that collects **all** services and resources (**ELE**).
- **Next steps**: develop and populate **ELG Release 2** with functional services and resources (early 2021); develop **ELG Release 3** (early 2022); establish the **ELG legal entity** (late 2021).



EUROPEAN LANGUAGE GRID

**Interested
to learn
more?**

- Visit our project website:

<https://www.european-language-grid.eu>

- Go to the live ELG:

<https://live.european-language-grid.eu>

- Like to have a user account for the grid? Send an email to:

contact@european-language-grid.eu



European Language Grid

Thank you!

Interested in participating?

Please get in touch:

contact@european-language-grid.eu

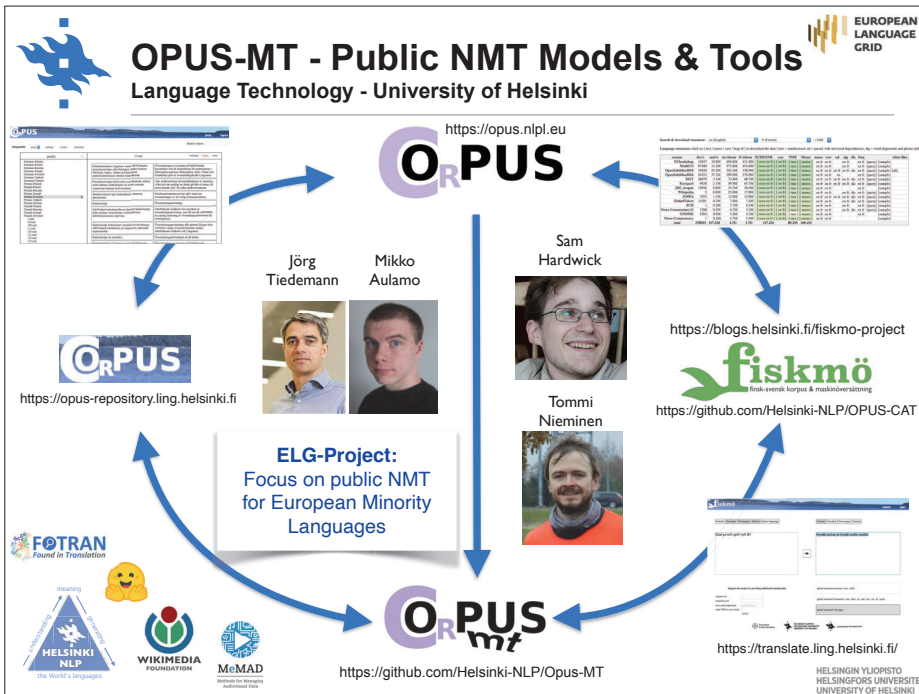


The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).

Katrin Marheinecke (DFKI, ELG Project Manager)
katrin.marheinecke@dfki.de

15-12-2020 – 4th Regional ELG Workshop – Finland
<http://www.european-language-grid.eu>

Coordinator: Dr. Georg Rehm (DFKI)



OPUS-MT
<https://github.com/Helsinki-NLP/Opus-MT>

Available software:

- MT server solution based on Marian NMT
- dockerised web-app, translation interface and API
- NMT training pipeline (OPUS-MT-train)
- CAT integration (OPUS-CAT)

Pre-trained translation models:

- number of bilingual models: 1,048
- number of multilingual models: 53
- number of supported source languages: 229
- number of supported target languages: 222
- number of supported language pairs: 1,715

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Prototype for Sámi language NMT
<https://translate.ling.helsinki.fi/ui/sami>

sami

Norwegian Bokmål Northern Sami

Grønningen er et vernet boligområde på Notodden. Navnet kommer av at alle husene i begynnelsen var malt grønne. Området ble bebygget av Notodden Salpeterfabriker (Norsk Hydro) med arbeiderboliger i perioden 1906–1911, og inneholder 25 vertikaldelte tomannsboliger og tre eneboliger. Hydros egen arkitektavdeling.

Grønningen lea suddjen ássanguovlu Notoddenis. Namma boahát das, ahte buot viesut álggos ledje ruonát. Guovllu huksejedje Notodden Salpeterfabriker (Norsk Hydro) ja bargiidviesut ledje áigodagas 1906–1911, ja sistisdoallá 25 ceakkojuohkijuvvon guovtteolbmoviesu ja eallhmu oktoásođaga. Hydros iežas.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

LANGUAGE TECHNOLOGY

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Embedded OPUS-MT in CAT tools
<https://github.com/Helsinki-NLP/OPUS-CAT>

SDL Trados Studio - Project 2

File Home Review Advanced View Add-Ins Help

Project Settings Batch Copy Paste File Actions Formatting QuickInsert Translation Memory Segment Actions Navigation

Editor

fiskmolest.bt.sdxiff [Translation]

1 Republiken president utnämnde Finland 75:e regering torsdagen den 6 juni.

2 Samtidigt befräddes presidenten ministrarna i Juha från medlemskap I den nya regering statsminister Antti sammanlagt 19 m Socialdemokrateri ministrar, Centern Vänsterförbundet folkpartiet två. Vid sitt konstituerade torsdagen den 6 j statsrådet om min arbetsfördelning.

tasavallan presidentti nimitti torstaina 6. kesäkuuta Suomen 75. hallituksen.

Translation Results - Fiskmo

Project Settings

Republiken president utnämnde Finland 75:e regering torsdagen den 6 juni.

Republiken president utnämnde Finland 75:e regering torsdagen den 6 juni.

tasavallan presidentti nimitti torstaina 6. kesäkuuta Suomen 75. hallituksen.

Fiskmo

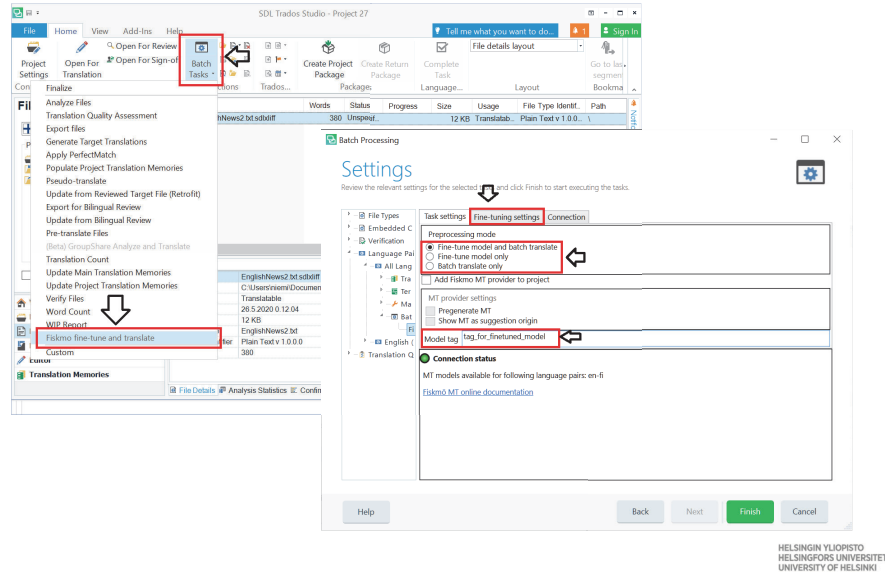
Translati... Fragme... Concord... Comm... TQAs (0) Messag...

0,00% Chars: 77 0:001



Embedded OPUS-MT in CAT tools

<https://github.com/Helsinki-NLP/OPUS-CAT>



Installation of the Tornado-based Web-App

Download the latest version from github:

```
git clone https://github.com/Helsinki-NLP/Opus-MT.git
```

Option 1: Manual setup

Install Marian MT. Follow the documentation at <https://marian-nmt.github.io/docs/>. After the installation, marian-server is expected to be present in path. If not place it in `/usr/local/bin`.

Install pre-requisites. Using a virtual environment is recommended.

```
pip install -r requirements.txt
```

Download the translation models from <https://github.com/Helsinki-NLP/Opus-MT/tree/master/models> and place it in models directory.

Then edit the services.json to point to that models.

And start the webserver.

```
python server.py
```

By default, it will use port 8888. Launch your browser to localhost:8888 to get the web interface. The languages configured in services.json will be available.

Option 2: Using Docker

```
docker-compose up
```

And launch your browser to localhost:8888

Tornado web-app server

Configuration

The server.py program accepts a configuration file in json format. By default it tries to use `config.json` in the current directory. But you can give a custom one using `-c` flag.

An example configuration file looks like this:

```
{
  "en": {
    "es": {
      "configuration": "./models/en-es/decoder.yml",
      "host": "localhost",
      "port": "10001"
    },
    "fi": {
      "configuration": "./models/en-fi/decoder.yml",
      "host": "localhost",
      "port": "10002"
    }
  }
}
```

This example configuration can provide MT service for en->es and en->fi language pairs.

- `configuration` points to a yaml file containing the decoder configuration usable by `marian-server`. If this value is not provided, Opus-MT will assume that the service is already running in a remote host and port as given in other options. If value is provided a new sub process will be created using `marian-server`.
- `host`: The host where the server is running.
- `port`: The port to be listened for `marian-server`.

Websocket client: Token-aligned output

```
echo "Mitä kuuluu? Käännös on hauskaa." | ./opusMT-client.py -H localhost -P 20000 -s fi -t en
```

This should return something like

```
{
  "alignment": [
    "0-0 0-2 1-1 2-3",
    "0-0 1-1 3-2 4-3 5-4"
  ],
  "result": "How are you? The translation is fun.",
  "server": "192.168.1.18:20001",
  "source": "fi",
  "source-segments": [
    "Mit\u00e4 kuuluu ?",
    "K\u00e4\u00e4nn\u00f6s on hauskaa ."
  ],
  "source-sentences": [
    "Mit\u00e4 kuuluu?",
    "K\u00e4\u00e4nn\u00f6s on hauskaa ."
  ],
  "target": "en",
  "target-segments": [
    "How are you ?",
    "The translation is fun ."
  ],
  "target-sentences": [
    "How are you?",
    "The translation is fun."
  ]
}
```


→   <https://github.com/Helsinki-NLP/Opus-MT-train>

Train Opus-MT models

This package includes scripts for training NMT models using MarianNMT and OPUS data for [OPUS-MT](#). More details are given in the [Makefile](#) but documentation needs to be improved. Also, the targets require a specific environment and right now only work well on the CSC HPC cluster in Finland.

Pre-trained models

The subdirectory `models` contains information about pre-trained models that can be downloaded from this project. They are distributed with a [CC-BY 4.0 license](#).

Quickstart

Setting up:

```
git clone https://github.com/Helsinki-NLP/OPUS-MT-train.git
git submodule update --init --recursive --remote
make install
```

Training a multilingual NMT model (Finnish and Estonian to Danish, Swedish and English):

```
make SRCLANGS="fi et" TRGLANGS="da sv en" train
make SRCLANGS="fi et" TRGLANGS="da sv en" eval
make SRCLANGS="fi et" TRGLANGS="da sv en" release
```



EUROPEAN
LANGUAGE
GRID

OPUS_{mt}

<https://github.com/Helsinki-NLP/Opus-MT>

```
{
  "alignment": {
    "0-0 0-1 1-1 2-3",
    "0-0 1-1 3-2 4-3 5-4"
  },
  "result": "How are you? The translation is fun.",
  "server": "192.168.1.18:20001",
  "source": "fi",
  "source-segments": [
    "Mit\u00e4 kuuluu ?",
    "K\u00e4\u00e4nn\u00e4ks n\u00e4n\u00e4 on hauskaa ."
  ],
  "source-sentences": [
    "Mit\u00e4 kuuluu?",
    "K\u00e4\u00e4nn\u00e4ks n\u00e4n\u00e4 on hauskaa."
  ],
  "target": "en",
  "target-segments": [
    "How are you?",
    "The translation is fun ."
  ],
  "target-sentences": [
    "How are you?",
    "The translation is fun."
  ]
}
```

Development demo:
<https://translate.ling.helsinki.fi/ui/sami>



OPUS-MT at huggingface
<https://huggingface.co/Helsinki-NLP>



HUGGING FACE

Back to all models

Model: Helsinki-NLP/opus-mt-ROMANCE-en

python rust marian bn-head seq2seq mx mt translation

Hosted inference API

fy translation

Your sentence here...

Compute

This model can be loaded on the inference API on demand.

Monthly model downloads

Helsinki-NLP/opus-mt-ROMANCE-en

31,680 downloads

per month

per week

per day

per hour

per minute

per second

per millisecond

per microsecond

per nanosecond

per picosecond

per femtosecond

per attosecond

per zeptosecond

per yoctosecond

per rontosecond

per attosecond

per femtosecond

per picosecond

per nanosecond

per microsecond

per millisecond

per second

per minute

per hour

per day

per week

per month

per year

per decade

per century

per millennium

per billion years

per trillion years

per quadrillion years

per quintillion years

per sextillion years

per septillion years

per octillion years

per nonillion years

per decillion years

per undecillion years

per duodecillion years

per tredecillion years

per quattuordecillion years

per quindecillion years

per sexdecillion years

per septendecillion years

per octodecillion years

per novecentillion years

per millinillion years

per billionillion years

per trillionillion years

per quadrillionillion years

per quintillionillion years

per sextillionillion years

per septillionillion years

per octillionillion years

per nonillionillion years

per decillionillion years

per undecillionillion years

per duodecillionillion years

per tredecillionillion years

per quattuordecillionillion years

per quindecillionillion years

per sexdecillionillion years

per septendecillionillion years

per octodecillionillion years

per novecentillionillion years

per millinillionillion years

per billionillionillion years

per trillionillionillion years

per quadrillionillionillion years

per quintillionillionillion years

per sextillionillionillion years

per septillionillionillion years

per octillionillionillion years

per nonillionillionillion years

per decillionillionillion years

per undecillionillionillion years

per duodecillionillionillion years

per tredecillionillionillion years

per quattuordecillionillionillion years

per quindecillionillionillion years

per sexdecillionillionillion years

per septendecillionillionillion years

per octodecillionillionillion years

per novecentillionillionillion years

per millinillionillionillion years

per billionillionillionillion years

per trillionillionillionillion years

per quadrillionillionillionillion years

per quintillionillionillionillion years

per sextillionillionillionillion years

per septillionillionillionillion years

per octillionillionillionillion years

per nonillionillionillionillion years

per decillionillionillionillion years

per undecillionillionillionillion years

per duodecillionillionillionillion years

per tredecillionillionillionillion years

per quattuordecillionillionillionillion years

per quindecillionillionillionillion years

per sexdecillionillionillionillion years

per septendecillionillionillionillion years

per octodecillionillionillionillion years

per novecentillionillionillionillion years

per millinillionillionillionillion years

per billionillionillionillionillion years

per trillionillionillionillionillion years

per quadrillionillionillionillionillion years

per quintillionillionillionillionillion years

per sextillionillionillionillionillion years

per septillionillionillionillionillion years

per octillionillionillionillionillion years

per nonillionillionillionillionillion years

per decillionillionillionillionillion years

per undecillionillionillionillionillion years

per duodecillionillionillionillionillion years

per tredecillionillionillionillionillion years

per quattuordecillionillionillionillionillion years

per quindecillionillionillionillionillion years

per sexdecillionillionillionillionillion years

per septendecillionillionillionillionillion years

per octodecillionillionillionillionillion years

per novecentillionillionillionillionillion years

per millinillionillionillionillionillion years

per billionillionillionillionillionillion years

per trillionillionillionillionillionillion years

per quadrillionillionillionillionillionillion years

per quintillionillionillionillionillionillion years

per sextillionillionillionillionillionillion years

per septillionillionillionillionillionillion years

per octillionillionillionillionillionillion years

per nonillionillionillionillionillionillion years

per decillionillionillionillionillionillion years

per undecillionillionillionillionillionillion years

per duodecillionillionillionillionillionillion years

per tredecillionillionillionillionillionillion years

per quattuordecillionillionillionillionillionillion years

per quindecillionillionillionillionillionillion years

per sexdecillionillionillionillionillionillion years

per septendecillionillionillionillionillionillion years

per octodecillionillionillionillionillionillion years

per novecentillionillionillionillionillionillion years

per millinillionillionillionillionillionillion years

per billionillionillionillionillionillionillion years

per trillionillionillionillionillionillionillion years

per quadrillionillionillionillionillionillionillion years

per quintillionillionillionillionillionillionillion years

per sextillionillionillionillionillionillionillion years

per septillionillionillionillionillionillionillion years

per octillionillionillionillionillionillionillion years

per nonillionillionillionillionillionillionillion years

per decillionillionillionillionillionillionillion years

per undecillionillionillionillionillionillionillion years

per duodecillionillionillionillionillionillionillion years

per tredecillionillionillionillionillionillionillion years

per quattuordecillionillionillionillionillionillionillion years

per quindecillionillionillionillionillionillionillion years

per sexdecillionillionillionillionillionillionillion years

per septendecillionillionillionillionillionillionillion years

per octodecillionillionillionillionillionillionillion years

per novecentillionillionillionillionillionillionillion years

per millinillionillionillionillionillionillionillion years

per billionillionillionillionillionillionillionillion years

per trillionillionillionillionillionillionillionillion years

per quadrillionillionillionillionillionillionillionillion years

per quintillionillionillionillionillionillionillionillion years

per sextillionillionillionillionillionillionillionillion years

per septillionillionillionillionillionillionillionillion years

per octillionillionillionillionillionillionillionillion years

per nonillionillionillionillionillionillionillionillion years

per decillionillionillionillionillionillionillionillion years

per undecillionillionillionillionillionillionillionillion years

per duodecillionillionillionillionillionillionillionillion years

per tredecillionillionillionillionillionillionillionillion years

per quattuordecillionillionillionillionillionillionillionillion years

per quindecillionillionillionillionillionillionillionillion years

per sexdecillionillionillionillionillionillionillionillion years

per septendecillionillionillionillionillionillionillionillion years

per octodecillionillionillionillionillionillionillionillion years

per novecentillionillionillionillionillionillionillionillion years

per millinillionillionillionillionillionillionillionillion years

per billionillionillionillionillionillionillionillionillion years

per trillionillionillionillionillionillionillionillionillion years

per quadrillionillionillionillionillionillionillionillionillion years

per quintillionillionillionillionillionillionillionillionillion years

per sextillionillionillionillionillionillionillionillionillion years

per septillionillionillionillionillionillionillionillionillion years

per octillionillionillionillionillionillionillionillionillion years

per nonillionillionillionillionillionillionillionillionillion years

per decillionillionillionillionillionillionillionillionillion years

per undecillionillionillionillionillionillionillionillionillion years

per duodecillionillionillionillionillionillionillionillionillion years

per tredecillionillionillionillionillionillionillionillionillion years

per quattuordecillionillionillionillionillionillionillionillionillion years

per quindecillionillionillionillionillionillionillionillionillion years

per sexdecillionillionillionillionillionillionillionillionillion years

per septendecillionillionillionillionillionillionillionillionillion years

per octodecillionillionillionillionillionillionillionillionillion years

per novecentillionillionillionillionillionillionillionillionillion years

per millinillionillionillionillionillionillionillionillionillion years

per billionillionillionillionillionillionillionillionillionillion years

per trillionillionillionillionillionillionillionillionillionillion years

per quadrillionillionillionillionillionillionillionillionillionillion years

per quintillionillionillionillionillionillionillionillionillionillion years

per sextillionillionillionillionillionillionillionillionillionillion years

per septillionillionillionillionillionillionillionillionillionillion years

per octillionillionillionillionillionillionillionillionillionillion years

per nonillionillionillionillionillionillionillionillionillionillion years

per decillionillionillionillionillionillionillionillionillionillion years

per undecillionillionillionillionillionillionillionillionillionillion years

per duodecillionillionillionillionillionillionillionillionillionillion years

per tredecillionillionillionillionillionillionillionillionillionillion years

per quattuordecillionillionillionillionillionillionillionillionillionillion years

per quindecillionillionillionillionillionillionillionillionillionillion years

per sexdecillionillionillionillionillionillionillionillionillionillion years

per septendecillionillionillionillionillionillionillionillionillionillion years

per octodecillionillionillionillionillionillionillionillionillionillion years

per novecentillionillionillionillionillionillionillionillionillionillion years

per millinillionillionillionillionillionillionillionillionillionillion years

per billionillionillionillionillionillionillionillionillionillionillion years

per trillionillionillionillionillionillionillionillionillionillionillion years

per quadrillionillionillionillionillionillionillionillionillionillionillion years

per quintillionillionillionillionillionillionillionillionillionillionillion years

per sextillionillionillionillionillionillionillionillionillionillionillion years

per septillionillionillionillionillionillionillionillionillionillionillion years

per octillionillionillionillionillionillionillionillionillionillionillion years

per nonillionillionillionillionillionillionillionillionillionillionillion years

per decillionillionillionillionillionillionillionillionillionillionillion years

per undecillionillionillionillionillionillionillionillionillionillionillion years

per duodecillionillionillionillionillionillionillionillionillionillionillion years

per tredecillionillionillionillionillionillionillionillionillionillionillion years

per quattuordecillionillionillionillionillionillionillionillionillionillionillion years

per quindecillionillionillionillionillionillionillionillionillionillionillion years

per sexdecillionillionillionillionillionillionillionillionillionillionillion years

per septendecillionillionillionillionillionillionillionillionillionillionillion years

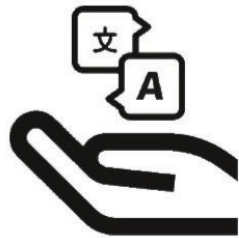
per octodecillionillionillionillionillionillionillionillionillionillionillion years

per novecentillionillionillionillionillionillionillionillionillionillionillion years

per millinillionillionillionillionillionillionillionillionillionillionillion years

per billionillionillionillionillionillionillionillionillionillionillionillion years

per trillionillion



**Provider of language services
and language technology
solutions**



**30+ years experience of NLP
development for text and
speech**



**Large collaboration network
with leading universities in the
Nordics**



LSDISCO - Lingsoft Solutions as Distributable Containers

- We will provide access to our speech and language technology tools through ELG to organisations and companies with language needs for the Nordic languages.
- We will make our solutions easily distributable with scalable integration in other environments and solutions
- We will make them available both for commercial and non-commercial usage

A photograph of a person carrying a young child on their shoulders, walking through a field of wildflowers. The scene is captured in a warm, golden light, suggesting a sunset or sunrise. The person is seen from behind, and the child is looking towards the camera. The field is filled with various types of wildflowers, and the background is a soft, out-of-focus landscape.

www.lingsoft.fi



Lingsoft Solutions

Text Analysis

Speech Recognition

Machine Translation

**Lemmatization, PoS
tagging**

Transcription

Customisation

NER and ontologies

Diarisation

Terminologies

**Spelling and
grammar**

Subtitling

**Professional CAT
support**

FI

SV

DA

NB

NN

EN

DE





Some Use Cases







Transcription & Text Editor

Lingsoft®

Speech recognition

Speech to Text




B **I** **U** |     **L** |  

Lingsoft on täyden palvelun kielitalo: Suomen suurin suomalainen käännöstoimisto ja Pohjoismaiden johtava kielipalvelujen ja kieliteknologiaratkaisujen toimittaja. Tarjoamme ketterät ja luotettavat käännöspalvelut, saavutettavat videotekstit tehokkaat tekstinkäsittelypalvelut sekä monipuoliset puheentunnistus- ja teks

käännöstoimisto

LMC Proofing

Ignore

 Paste Ctrl+V

Machine Translation for Professional Translators

Translation Results - none | Fragment Matches - none | Concordance Search | Comments | TQAs (0) | Messages

MTLAB_TEST_FI-EN.txt.sdlxiff [Translation]*

1	Mikä on Lingsoft MT Lab?		
2	Lingsoft MT Lab on Trados Studio -laajennus, jonka kautta konekäännös tarjotaan käyttöön Lingsoft Post Editing -toissää.		Lingsoft MT Lab is a Trados studio Post Editing.
3	Konekäännökset tuotetaan Lingsoftin omilla konekääntimillä, joita kehitetään jatkuvasti.		Machine translations are produced continuously developed.
4	Pikaohjeet		Quick Tips
5	Lataa Studio-versiotasi vastaava MT Lab -versio (https://mt.lingsoft.fi/download) ja asenna sdplugin-tiedosto.		Download the corresponding version of Studio and install the sdplugin.
6	Tuo MT Lab -ikkuna näkyviin Studion editorinäköymässä valitsemalla View-välilehden yläpalkista Lingsoft MT Lab (alpha).		Click Lingsoft MT Lab (Alpha) on the Editor view.
7	Konekäännös näytetään MT results -ikkunassa.		The machine translation is displayed in the MT results window.
8	Kun käännät, kopioi konekäännökset kohdesegmenttiin Alt+Shift+X-pikanäppäimellä ja viimeistele käännös niiden pohjalta.		When you turn, use the Alt + Shift + X target segment, and then use the target segment to finish the translation.
9	Kun työ on valmis, valitse Home-välilehden Lingsoft MT Lab -osasta Finalize Project.		When the job is finished, click the Finalize Project button in the Lingsoft MT Lab section of the Home tab.

MTLAB_TEST_FI-EN.txt

LingsoftMTLab (alpha)

MT results

What is Lingsoft MT Lab?

Metadata and Entity Recognition

Lingsoft Analyser Demo

Text	Analysis	CoNLL-U	JSON output	Unknowns	Terms	Ontology	Annotation
------	----------	---------	-------------	----------	-------	----------	------------

Lingsoft on täyden palvelun kielitalo: Suomen suurin suomalainen käännöstoimisto ja Pohjoismaiden johtava kielipalvelujen ja kieliteknologiaratkaisujen toimittaja. Tarjoan Suomi (LOC) ja luotettavat käännöspalvelut, saavutettavat videotekstitykset, tehokkaat tekstinkäsittelypalvelut sekä monipuoliset punnentunnistus- ja tekstianalytiikkaratkaisut.