

EUROPEAN LANGUAGE GRID



Large Language Models and European Language Equality – Where do we stand and what do we need to do?

Prof. Dr. Georg Rehm (DFKI) – Coordinator ELG, Co-Coordinator ELE

14-06-2022 Large language models: pre-training with a twist

<http://www.european-language-grid.eu> – <https://european-language-equality.eu> – <https://opengpt-x.de>

Language Models

- Language Models (LMs), often Large Language Models (LLMs), define the state of the art in various Natural Language Processing (NLP) and Language Technology (LT) tasks.
- For example, speech recognition, machine translation, natural language generation, part-of-speech tagging, parsing etc.
- LMs are probability distributions over sequences of words. LMs are neither sentient nor conscious.
- LMs are trained using vast amounts of monolingual or multilingual text data, nowadays sometimes even including visual data (photos). Technology: deep neural networks (Transformers).
- Pre-training LMs has very demanding requirements regarding compute infrastructure (some LMs have cost millions to train) – only a few organisations can train an LM from scratch.
- LMs can also be generative, i.e., they can generate text based on an input prompt.
- LMs can be fine-tuned to specific tasks (supervised machine learning) using rather small labelled datasets, for example, for question answering, sentiment analysis, next-sentence prediction etc.
- LMs are an incredibly dynamic research topic – *many* scientific papers, *many* open questions.

Global NLP/LT Market – Driven, to a large extent, by LLMs



The Global Natural Language Processing (NLP) Market was valued at USD 10.72 billion in 2020, and it is expected to be worth USD 48.46 billion by 2026, registering a CAGR of 26.84% during the forecast period (2021-2026). Due to the ongoing Covid-19 pandemic the market is witnessing growth in healthcare sector.

\$48.46B in 2026

Natural language processing market revenue worldwide 2017-2025

Published by [Statista Research Department](#), Mar 17, 2022



Worldwide revenue from the natural language processing (NLP) market is forecast to increase rapidly in the next few years. The NLP market is predicted to be almost 14 times larger in 2025 than it was in 2017, increasing from around three billion U.S. dollars in 2017 to over 43 billion in 2025. Natural language processing (NLP) is a branch of artificial intelligence (AI) that helps computers understand, interpret and manipulate human language. Drawing from computer science and computational linguistics among other

\$43B in 2025



MARKETSANDMARKETS

[HOME](#) [ABOUT US](#) [LEADERSHIP TEAM](#) [RESEARCH EXPERTS](#) [BRIEFINGS](#) [CAREERS](#) [CONTACT US](#)

[KNOWLEDGE STORE](#) [REPORTS](#) [RESEARCH INSIGHT](#) [RESOURCE CENTER](#) [EVENTS](#)

[271 Pages Report] The global Natural Language Processing market size to grow from USD 11.6 billion in 2020 to USD 35.1 billion by 2026, at a Compound Annual Growth Rate (CAGR) of 20.3% during the forecast period. Growing demand for cloud-based NLP solutions to reduce overall costs and better scalability and increasing usage of smart devices to facilitate smart environments are expected to drive the natural language processing market growth. The rise in the adoption of NLP-based applications across verticals to enhance customer experience and increase in investments in the healthcare vertical is expected to offer opportunities for NLP vendors.

\$35.1B in 2025

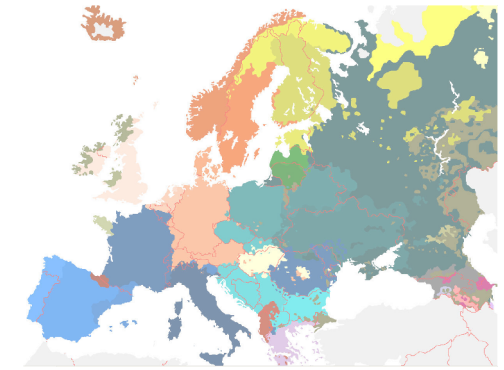
Outline

1. European Language Grid (ELG)
2. OpenGPT-X
3. European Language Equality (ELE)
4. Summary and Conclusions



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
- EP resolution (2018) asked for the development of a LT platform



European Parliament

2014-2019



TEXTS ADOPTED
Provisional edition

P8_TA-PROV(2018)0332

Language equality in the digital age

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))

The European Parliament,

- having regard to Articles 2 and 3(3) of the Treaty on the Functioning of the European Union (TFEU),
- having regard to Articles 21(1) and 22 of the Charter of Fundamental Rights of the European Union,
- having regard to the 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage,
- having regard to Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information¹,
- having regard to Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information²,
- having regard to Decision (EU) 2015/2240 of the European Parliament and of the Council of 25 November 2015 establishing a programme on interoperability solutions and common frameworks for European public administrations, businesses and citizens (ISA2 programme) as a means for modernising the public sector³,
- having regard to the Council resolution of 21 November 2008 on a European strategy for multilingualism (2008/C 320/01)⁴,
- having regard to the Council decision of 3 December 2013 establishing the specific programme implementing Horizon 2020 – the Framework Programme for Research and

¹ OJ L 345, 31.12.2003, p. 90.

² OJ L 175, 27.6.2013, p. 1.

³ OJ L 318, 4.12.2015, p. 1.

⁴ OJ C 320, 16.12.2008, p. 1.



META FORUM 2010

T4ME

T4ME

EU-funded project (Seventh Framework Programme) working on technologies for the Multilingual European Information Society

META-NET

META-NET

Established in 2010, META-NET is a network of Excellence consisting of 60 research centres from 34 countries, building the technological foundations of a multilingual European information society

META-FORUM 2010

"Challenges for Multilingual Europe" (November 17/18, 2010)



META SHARE

META FORUM 2015



CRACKER

EU-funded project CRACKER (Horizon2020) pushing towards an improvement of MT research in terms of efficiency and effectiveness (2015 – 2017)

Cracking the Language Barrier Federation

Founded in 2015, the federation has been assembling European research and innovation projects as well as all related community organisations working on multilingual technologies

META-FORUM 2015

Conference "Technologies for the Multilingual Digital Single Market" (Riga – April 27, 2015)

Strategic Agenda for the Multilingual Digital Single Market (Version 0.5)

Launch of the Strategic Agenda for the Multilingual Digital Single Market titled "Technologies for Overcoming Language Barriers towards a truly integrated European Online Market" (April 2015)

Riga Summit on the Multilingual Digital Single Market

Summit "Shape the future of the multilingual digital single market" (April 27–29, 2015)

META FORUM 2017

"Language Equality in the Digital Age"

Workshop on "Language Equality in the Digital Age", commissioned by the EU Parliament's Science and Technology Options Assessment Committee (STOA) (January 2017)

"Language equality in the digital age: Towards a Human Language Project"

Launch of the study on "Language equality in the digital age: Towards a Human Language Project", commissioned by the EU Parliament (March 2017)

META-FORUM 2017

Conference "Towards a Human Language Project" (Brussels – November 13/14, 2017)

Strategic Research and Innovation Agenda (V1.0)

Launch of the Strategic Research and Innovation Agenda titled "Language Technologies for Multilingual Europe – Towards a Human Language Project" (December 2017)



META FORUM 2012

META-FORUM 2012

Conference "A Strategy for Multilingual Europe" (Brussels – June 20/21, 2012)

META-NET White Papers

Release of 32 volumes on 31 languages, revealing that there is a severe threat of digital extinction for at least 21 European languages (December 2012)

2013

2014

"State of the Art of Machine Translation – Current Challenges and Future Opportunities" Workshop on "State of the Art of Machine Translation", commissioned by the EU Parliament (December 2013)

2014

Strategic Research and Innovation Agenda (Version 0.9)

Launch of the Strategic Research and Innovation Agenda titled "Language as a Data Type and Key Challenge for Big Data" (July 2016)

2016

ELG Final Proposal Submission

Final submission on Feb. 20, 2018

2018

2010

2011

2012

META-NORD

EU-funded (ICT PSP) Project to establish an open linguistic infrastructure in the Baltic and Nordic countries (2011 – 2013)

METANET4U

EU-funded project (ICT Policy Support Programme) to enhance the European Linguistic Infrastructure (2011 – 2013)

CESAR

EU-funded project (ICT Policy Support Programme) functioning as a part of META-NET to standardise language resources and tools (2011 – 2013)

META-FORUM 2011

Conference "Solutions for Multilingual Europe" (Budapest – June 27/28, 2011)

META FORUM 2011

META FORUM 2013

META FORUM 2016

META-SHARE

Initiated in 2013, META-SHARE has functioned as an open and secure network of repositories for sharing and exchanging language data, tools and services

Strategic Research Agenda for Multilingual Europe 2020

Launch of the Strategic Research Agenda for Multilingual Europe 2020 (January 2013)

META-FORUM 2013

Conference "Connecting Europe for New Horizons" (Berlin – September 19/20, 2013)

ELG consortium



EUROPEAN LANGUAGE GRID

2019-2022

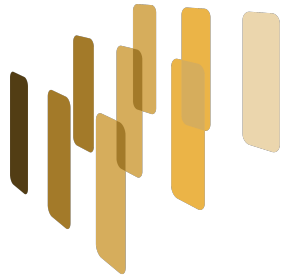


EUROPEAN LANGUAGE EQUALITY

2021-2022



European Language Grid



EUROPEAN LANGUAGE GRID

Objectives (Selection)

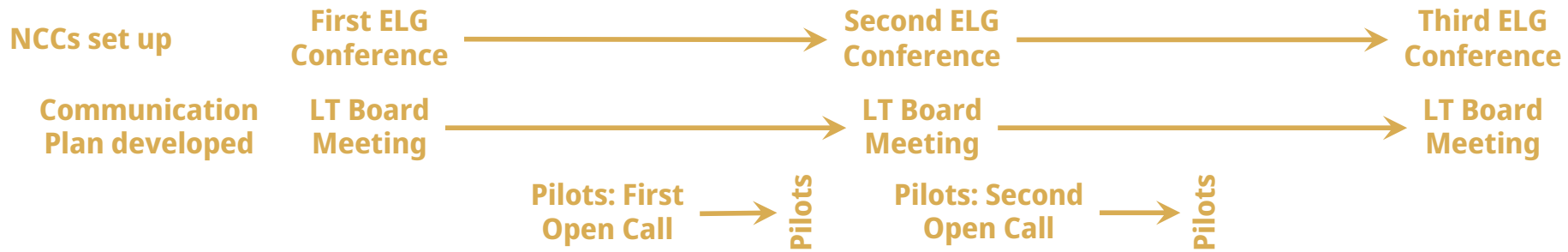
1. Establish the ELG as the primary Language Technology platform and market place in Europe to tackle the fragmentation of the European LT landscape.
2. ELG as a platform for commercial and non-commercial, industry-related LTs (functional and non-functional).
3. Enable the European LT community to upload services and data sets, 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.



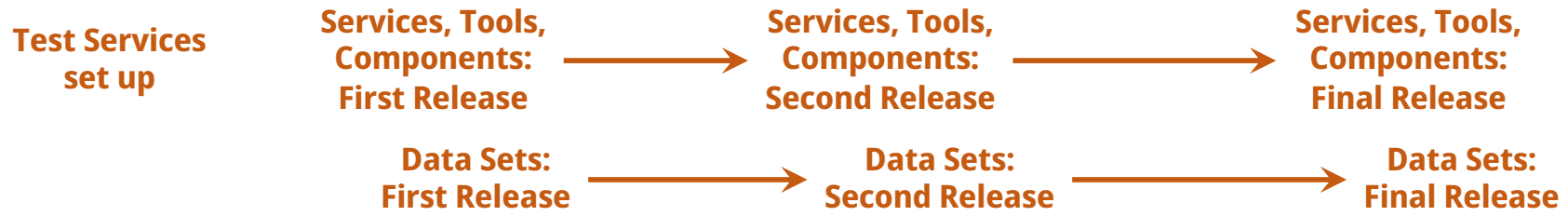
Kick-off meeting, 22/23 January 2019



Grid Community



Grid Content



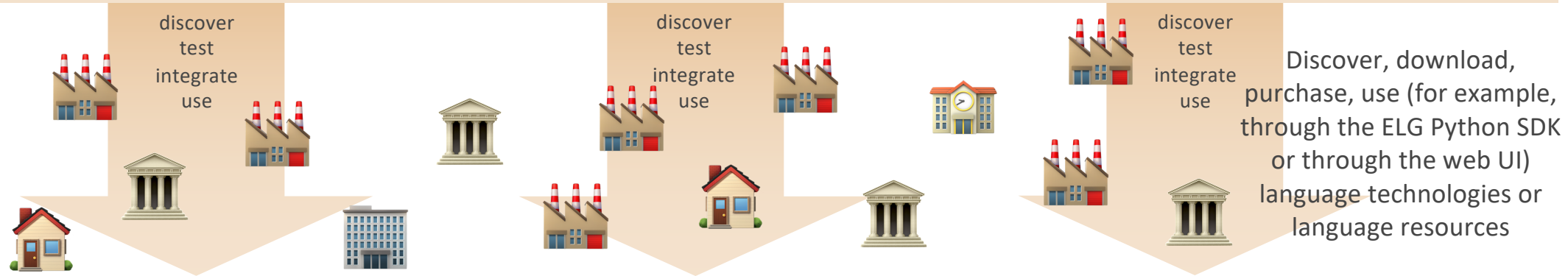
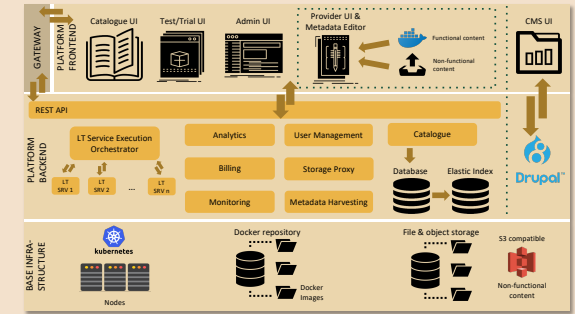
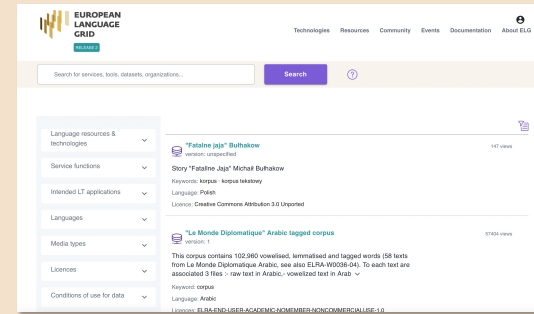
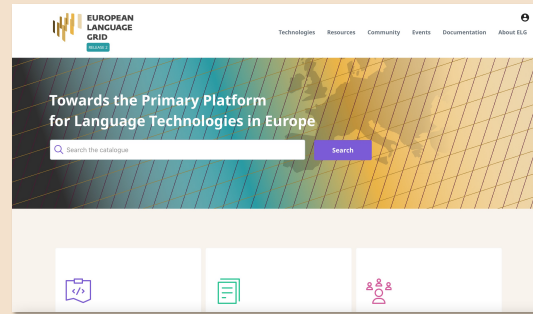
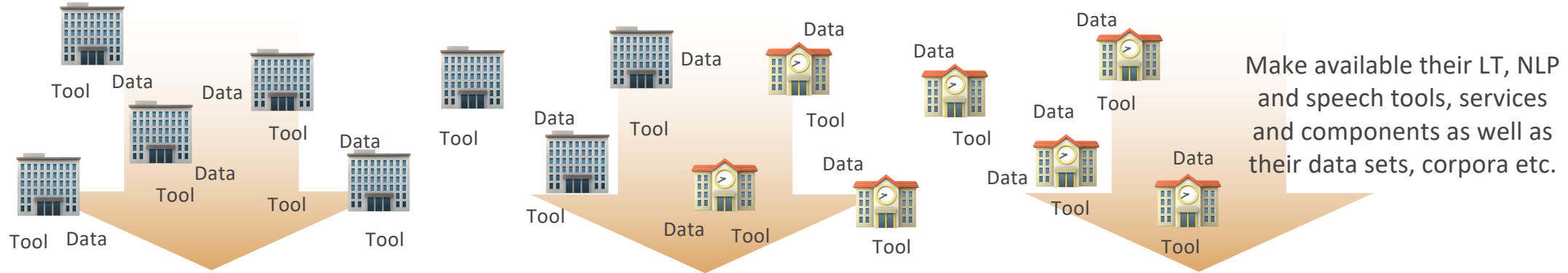
Grid Platform



Mgmt.



Developers of Language Technologies: Companies, Universities, Research Centres (approx. 1750-2000 organisations in total in Europe)



META-SHARE



harvest and integrate



ELRC-SHARE



harvest and integrate



OPUS



harvest and integrate



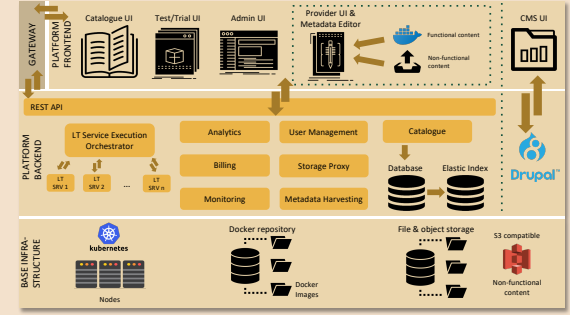
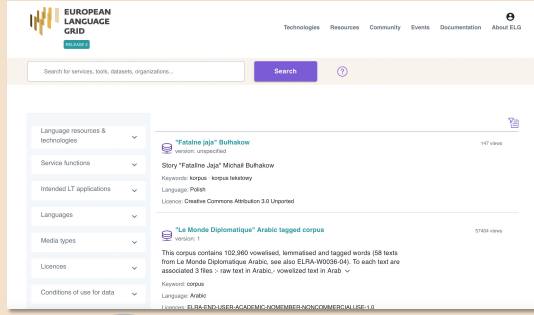
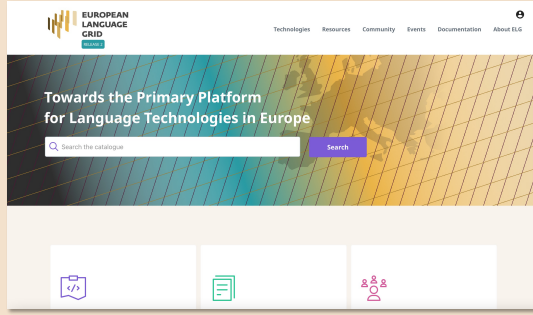
Zenodo



harvest and integrate



Harvest all relevant LT/LR repositories on a regular basis, i.e., collect metadata about resources and make them available through ELG for increased discovery and visibility



harvest and integrate



ELRA



harvest and integrate



HuggingFace



harvest and integrate



CLARIN



harvest and integrate



...



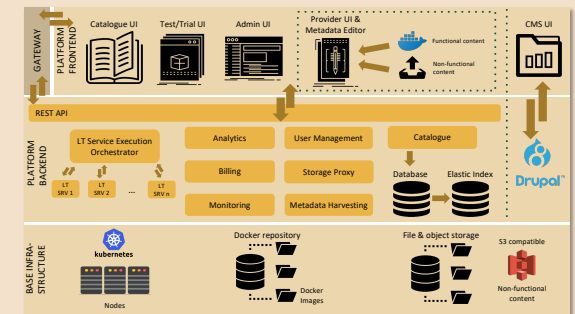
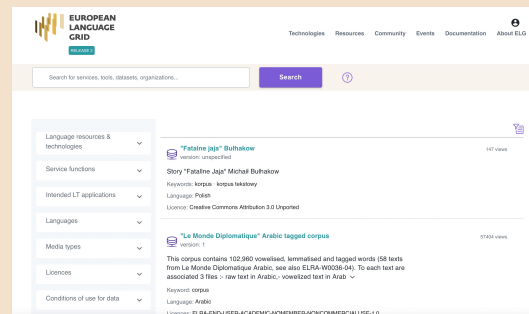
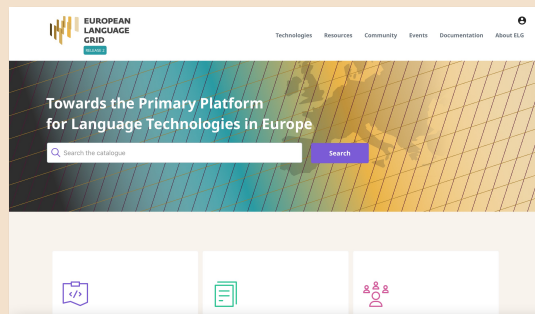
harvest and integrate



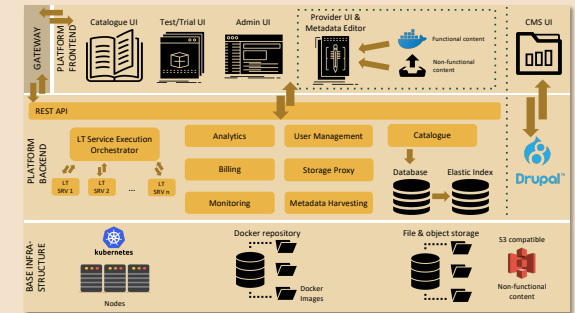
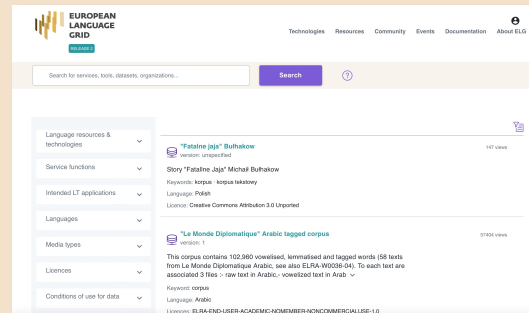
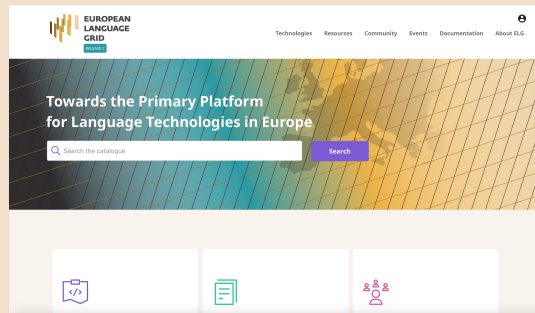
...



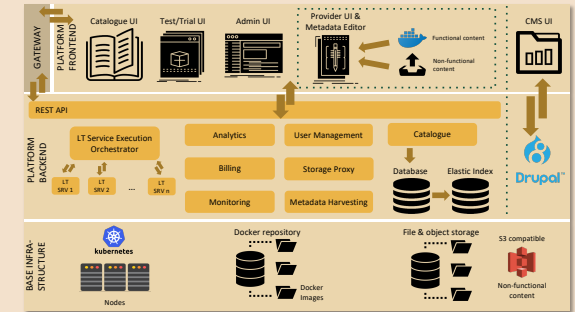
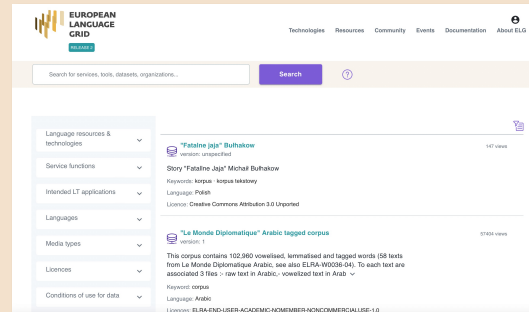
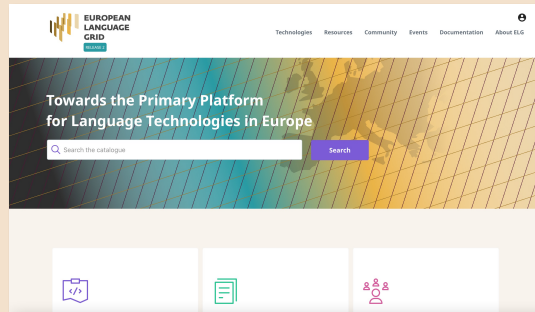
Harvesting of Language Technology-related Resource, Data, Tool and Model Repositories



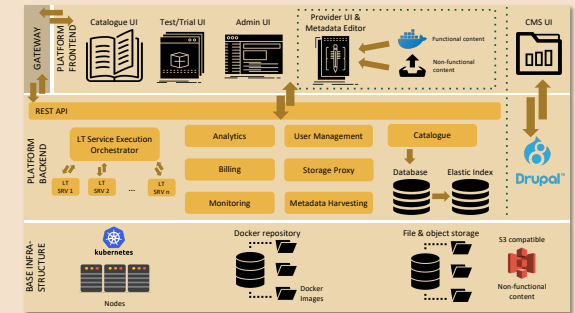
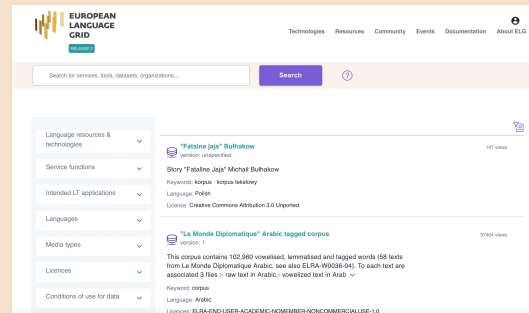
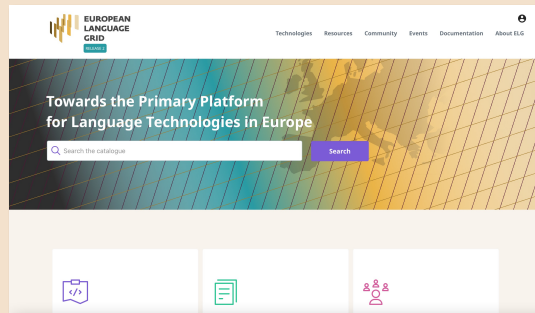
European Language Grid is a **joint technology platform** for the whole European Language Technology community (approx. 1750-2000 organisations)



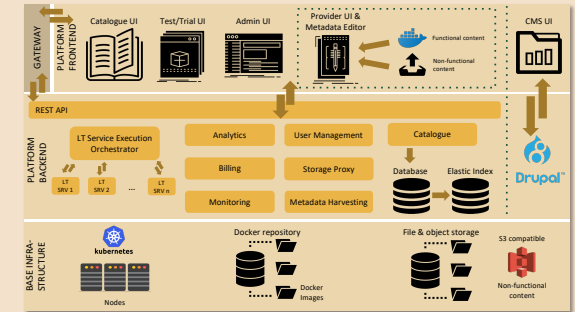
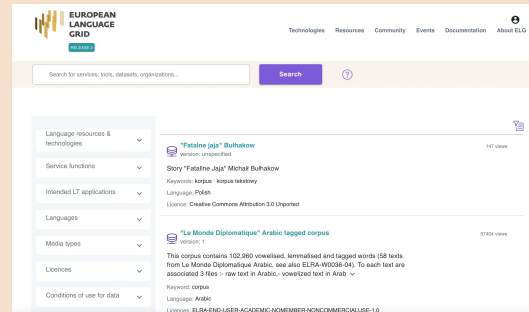
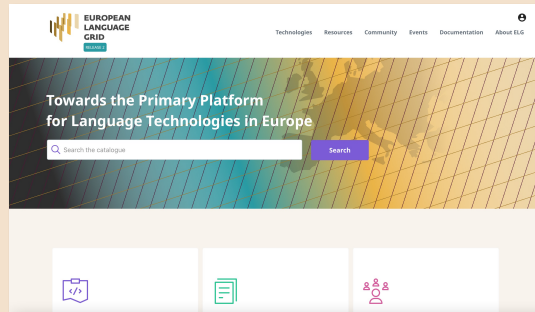
European Language Grid is a joint tool and resource sharing platform



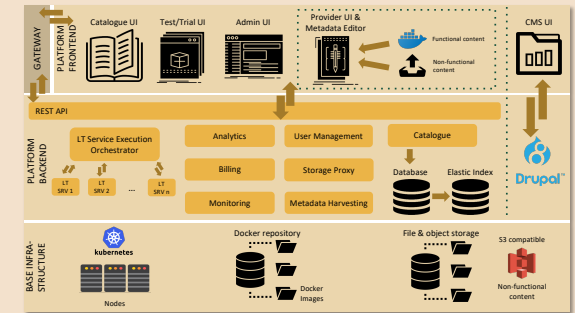
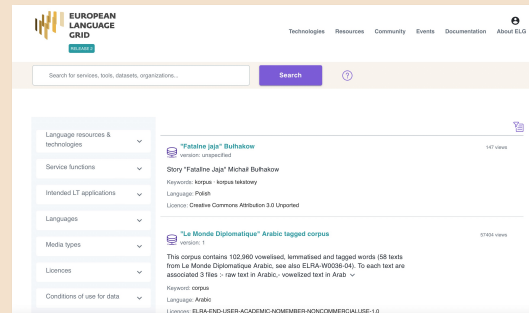
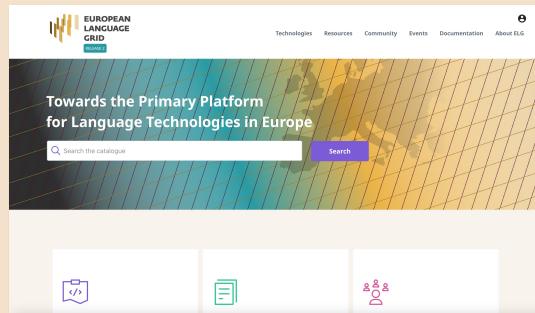
European Language Grid is a language data space



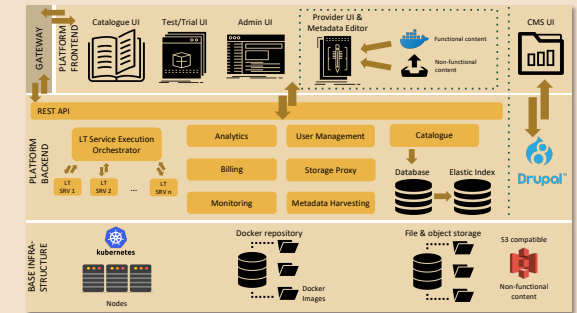
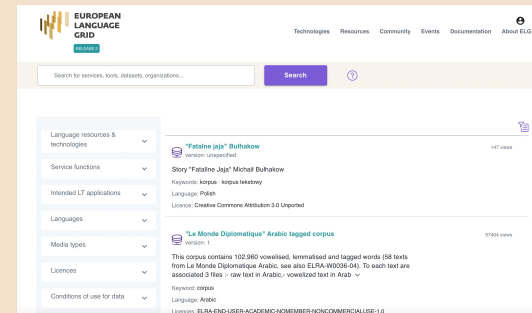
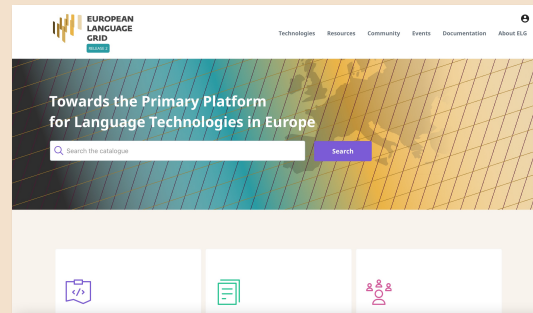
European Language Grid is a marketplace for the whole European Language Technology community



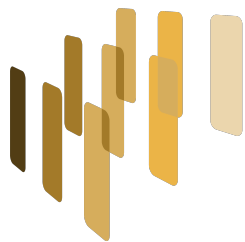
European Language Grid is the **community of European Language Technology developers**



European Language Grid is the **yellow pages** of the European Language Technology community



ELG makes available all European Language Technologies and Language Resources in a one-stop-shop.



EUROPEAN LANGUAGE GRID

ELG Release 3 (June 2022)

12,000+ Resources without organisations

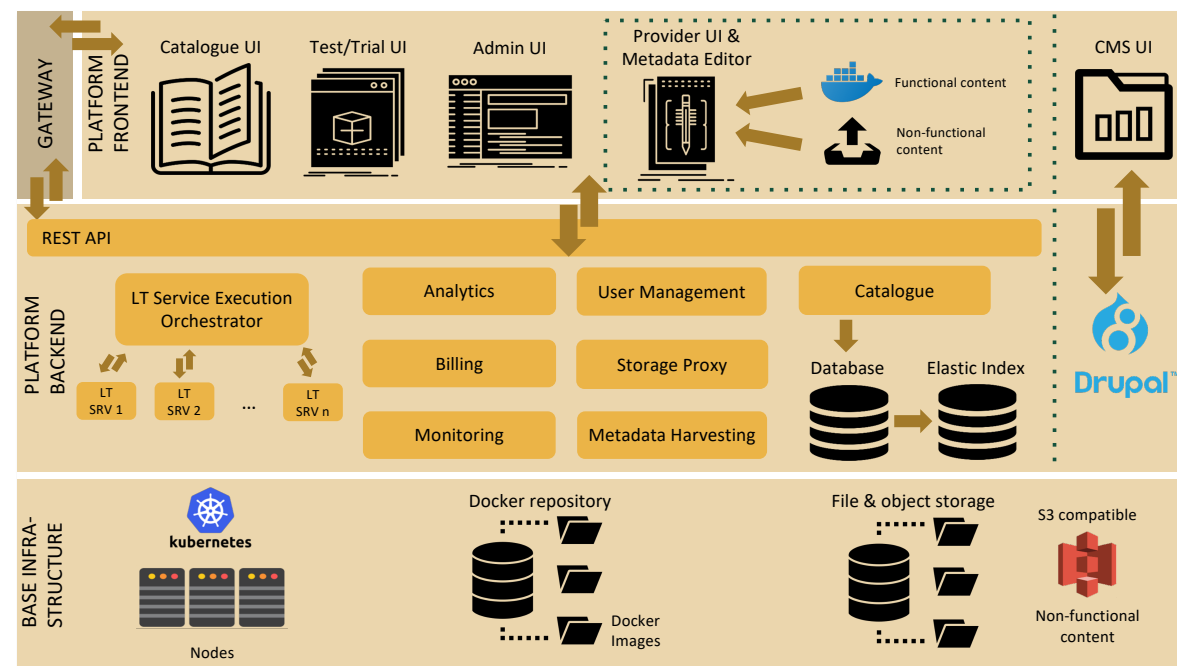
- 6259 corpora and data sets
- 3182 services and tools
 - 772 services available in and through ELG
- 2254 lexical/conceptual resources
- 465 models, grammars, lang. descriptions
- 1778 organisations (research orgs., companies)

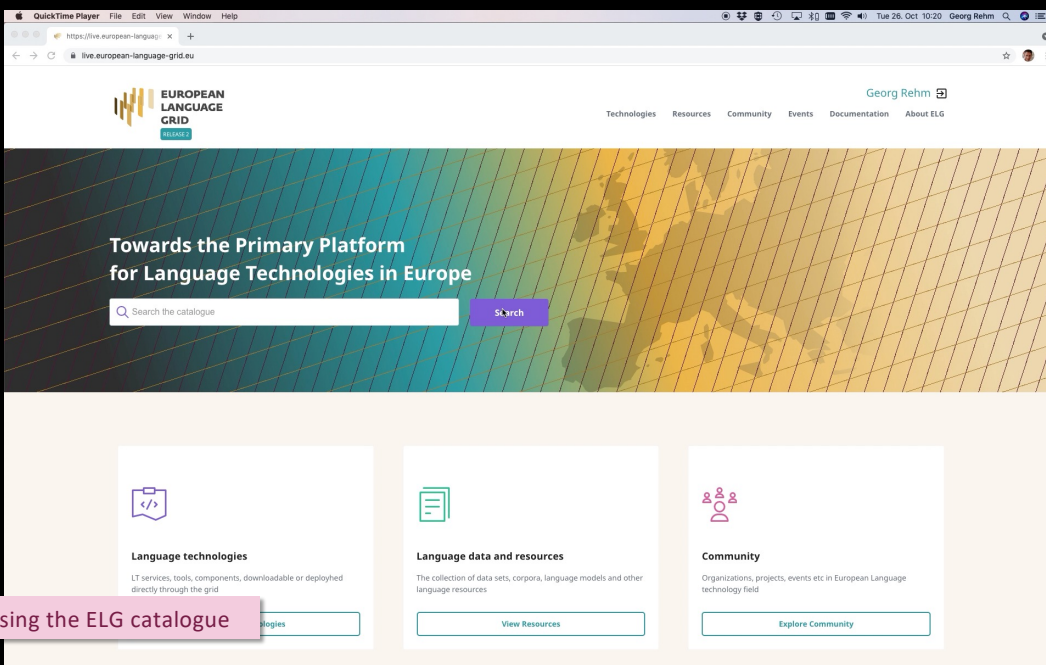
Users can connect to the ELG cloud platform via ELG APIs, remote APIs, ELG GUI, Python SDK, download of containers or source code.

+ ELG / ELE	(9032)
+ ELRC-SHARE	(1299)
+ ELRA Catalogue of Language Resources	(1180)
+ Zenodo	(513)
+ LINDAT/CLARIAH-CZ	(507)
+ Hugging Face	(385)

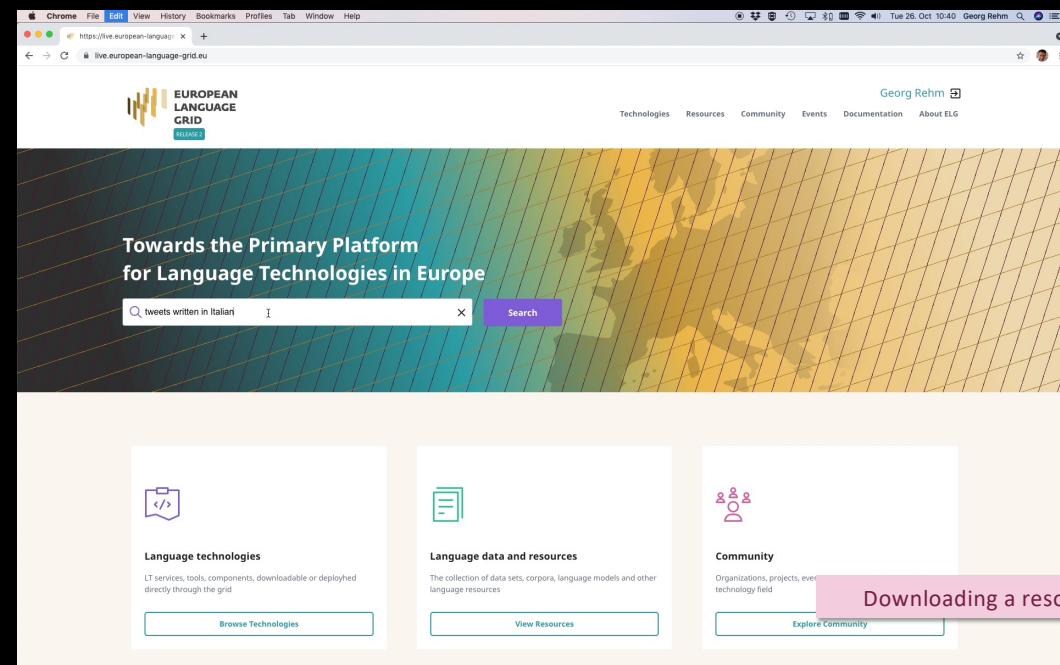
+ CLARIN-PL	(360)
+ Quantum Stat Datasets	(261)
+ CLARIN.SI	(218)
+ LREC Shared LRs (ELRA)	(144)
+ META-SHARE/ILSP	(69)
+ META-SHARE/DFKI	(2)
+ TUdatalib	(1)

Sources of data sets

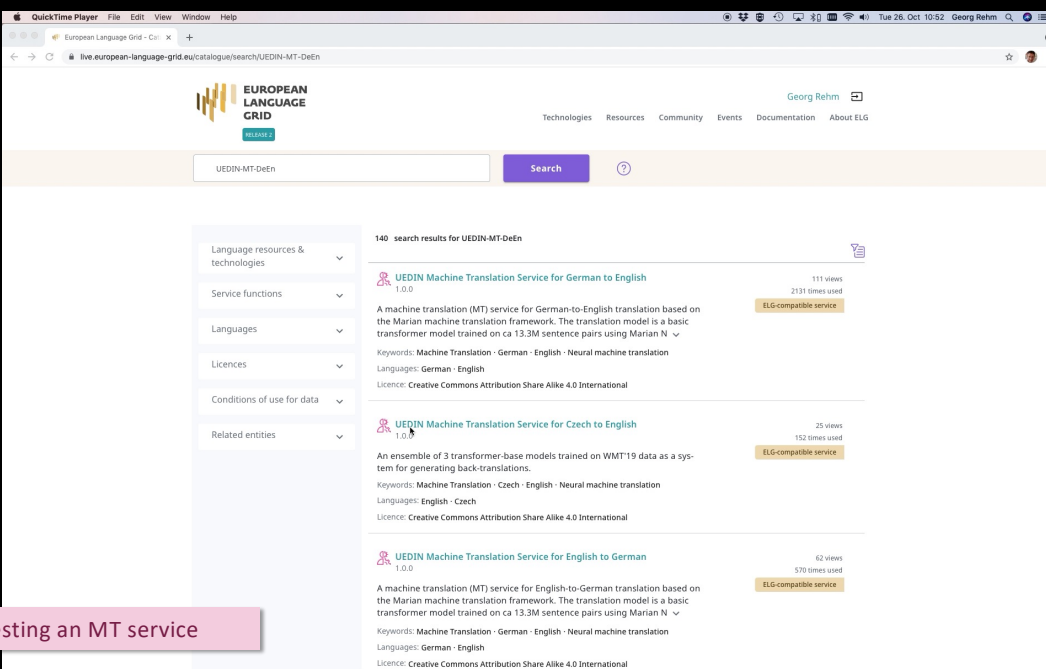




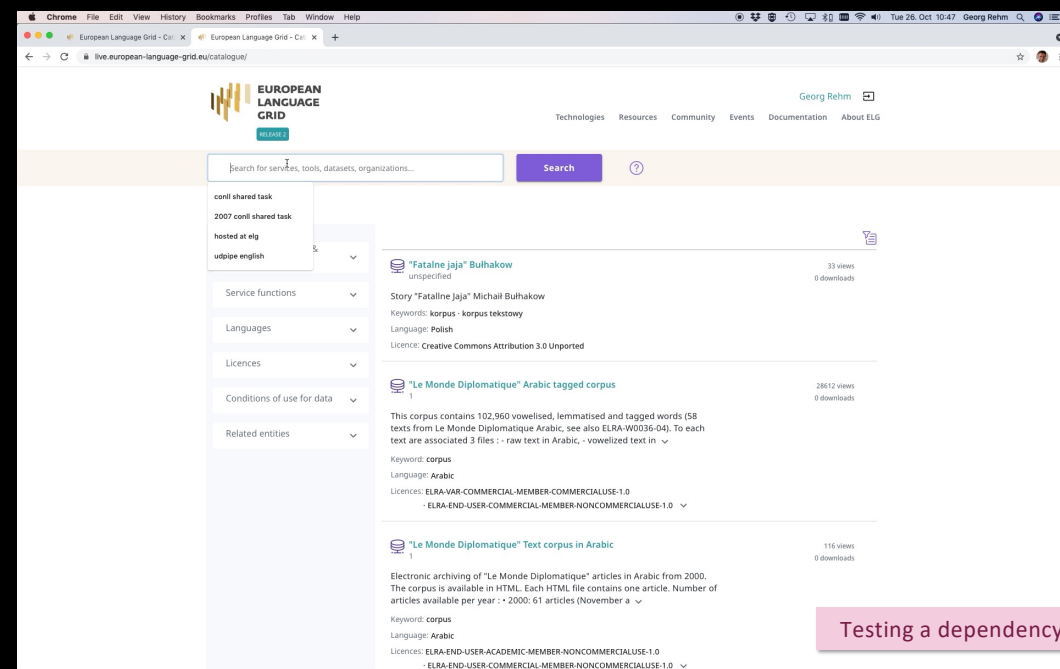
Browsing the ELG catalogue



Downloading a resource



Testing an MT service



Testing a dependency parser

Stakeholders and Users and Collaborators

32 ELG National Competence Centres

ELG Open Calls: 15 pilot projects

ELE – European Language Equality

AI4EU – European AI on demand platform

Data Spaces and Research Data Initiatives

ICT-29b research and innovation projects

Other related projects: CEF INEA projects (MAPA, NTEU, Microservices at your Service etc.), OpenGPT-X, NFDI4DS etc.

Additional initiatives: CLAIRE, AI PPP, ELRC, ECSPM, EFNIL, BDVA, CLARIN, W3C etc.



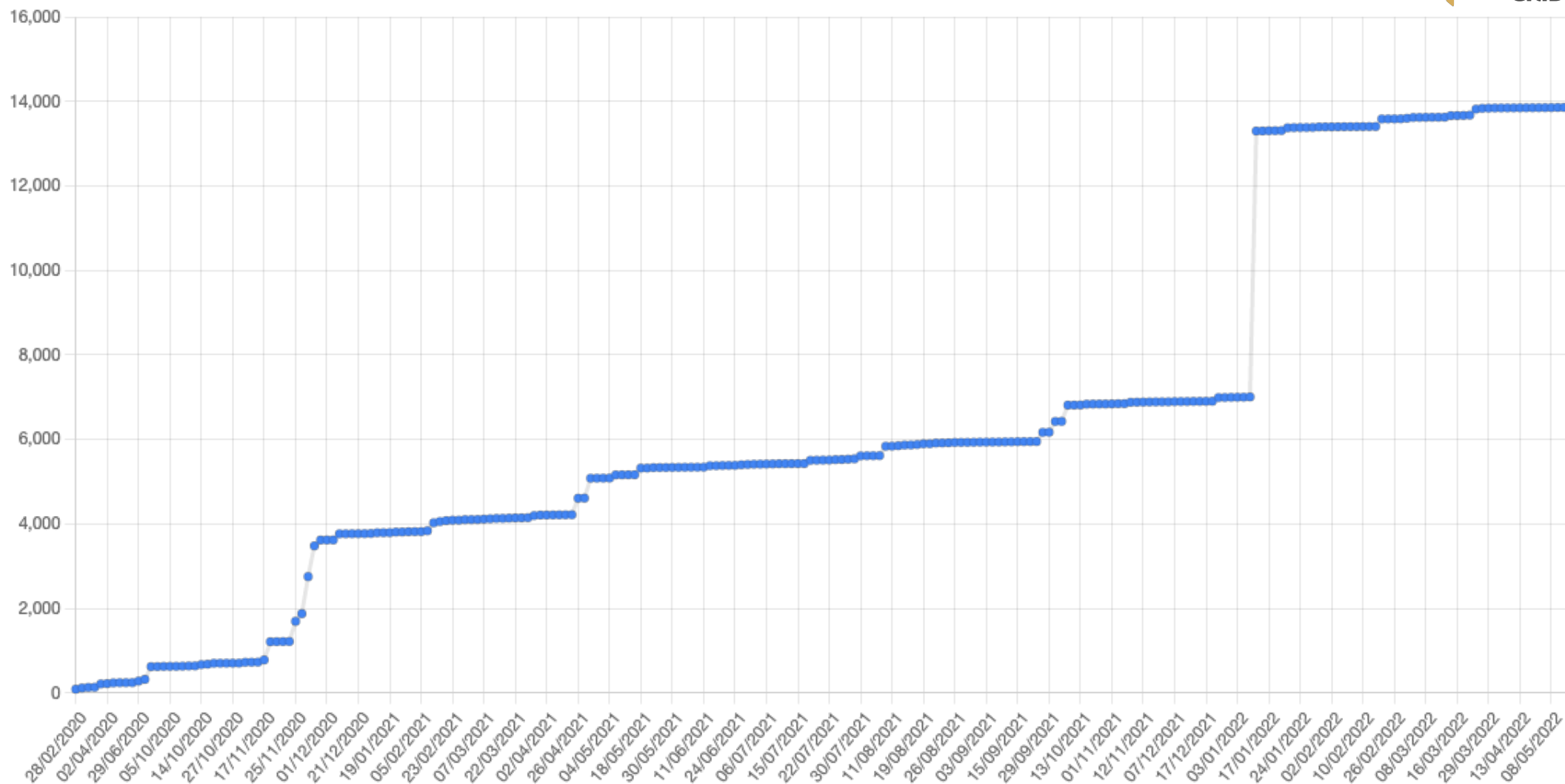
Global Under Resourced Media Translation



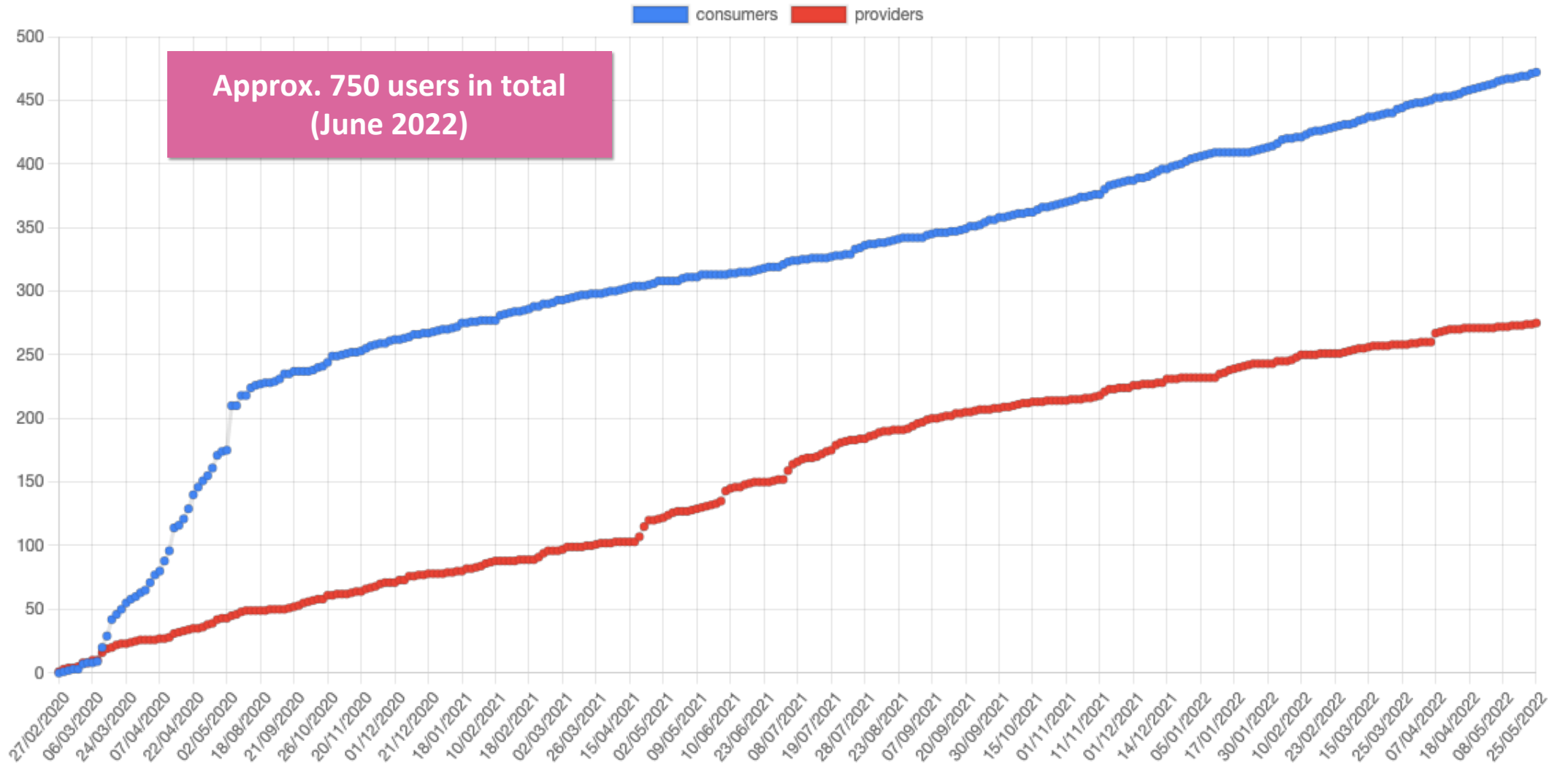
COMPRISE



Number of Resources over Time



Number of Users (Consumers of LT, Providers of LT)



Additional Information

META-FORUM 2022 – June 08-10, Brussels

Joining the European Language Grid

META-FORUM 2021 – November 15-17, virtual

Using the European Language Grid

META-FORUM 2020 – December 01-03, virtual

Piloting the European Language Grid

META-FORUM 2019 – October 08/09, Brussels

Introducing the European Language Grid

META-FORUM 2017 – November 13/14, Brussels

Towards a Human Language Project

META-FORUM 2016 – July 04/05, Lisbon

Beyond Multilingual Europe

META-FORUM 2015 – April 27, Riga

Technologies for the Multilingual Digital Single Market

META-FORUM 2013 – September 19/20, Berlin

Connecting Europe for New Horizons

META-FORUM 2012 – June 20/21, Brussels

A Strategy for Multilingual Europe

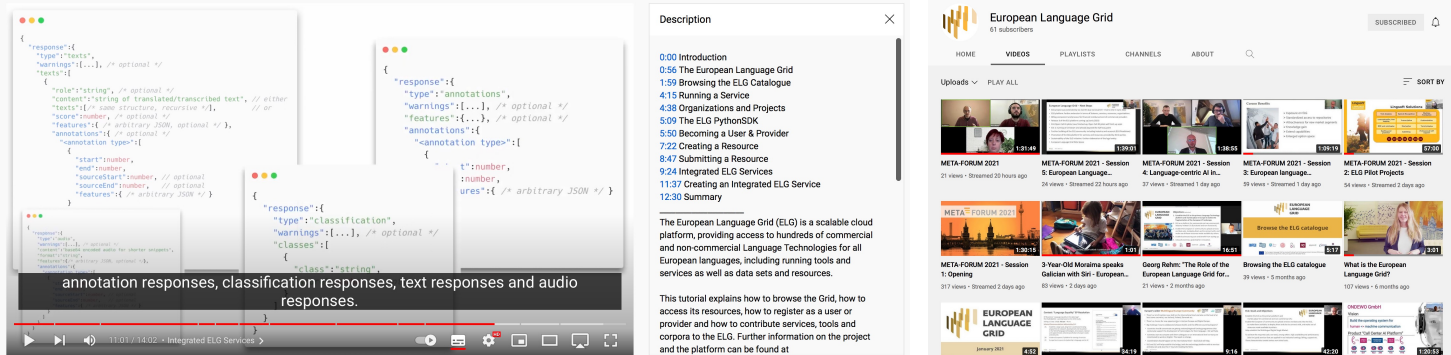
META-FORUM 2011 – June 27/28, Budapest

Solutions for Multilingual Europe

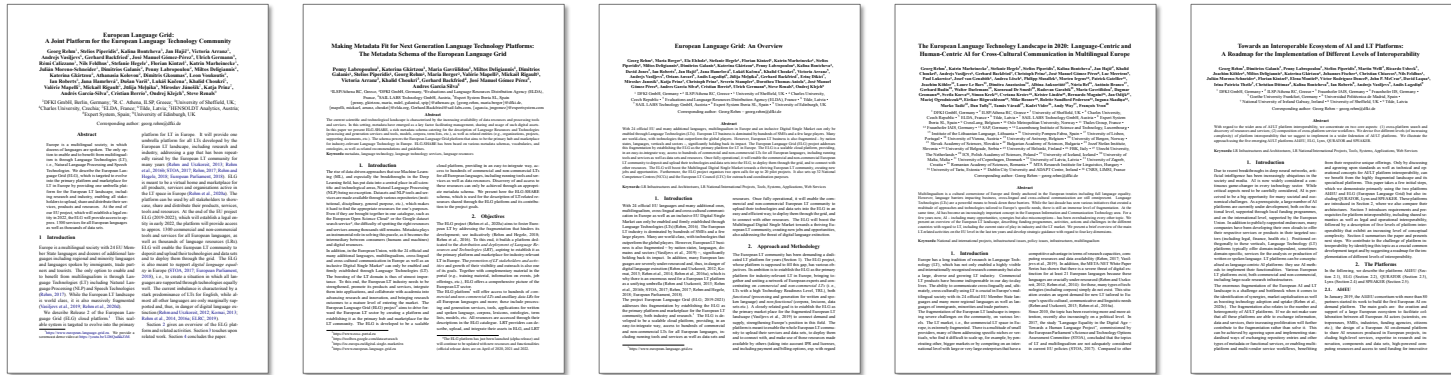
META-FORUM 2010 – November 17/18, Brussels

Challenges for Multilingual Europe

All META-FORUM 2020, 2021 and 2022 sessions and also various tutorial videos are available on the ELG YouTube channel:
<https://www.youtube.com/channel/UCarEHmsWT2JslcvvWkbhL4A>



The screenshot shows a video player with a code editor on the left displaying JSON code for annotations. The code includes fields like 'type', 'warnings', 'features', 'start', 'end', 'source', and 'classes'. The right side of the video shows a description of the ELG platform, mentioning its role as a scalable cloud platform for running tools and services, and providing access to hundreds of commercial and non-commercial language technologies.



This block contains a grid of document thumbnails. The first row includes 'A Joint Platform for the European Language Technology Community', 'Making Metadata for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid', and 'European Language Grid: An Overview'. The second row includes 'The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communications in Multilingual Europe' and 'Towards an Interoperable Ecosystem of AI and L1 Platform: A Roadmap for the Implementation of Different Levels of Interoperability'.

Technical Documentation: <https://european-language-grid.readthedocs.io>
Schema documentation: <https://gitlab.com/european-language-grid/platform/ELG-SHARE-schema>
Feedback: <https://gitlab.com/european-language-grid/platform/elg-platform>
Helpdesk: <https://www.european-language-grid.eu/contact/>

Outline

1. European Language Grid (ELG)
2. **OpenGPT-X**
3. European Language Equality (ELE)
4. Summary and Conclusions



OpenGPT-X secures European and German Digital Sovereignty in the field of AI



LLMs are almost exclusively developed by **US-American** and **Chinese** enterprises (e.g., GPT-3 and WuDao 2.0)

Access to LLMs for industry and research is often limited, as, for example, **GPT-3 is licensed by Microsoft**

To **foster innovation** as well as to **strengthen its ability to compete** there is a great demand for LLMs **“Made in Germany”**.

- Runtime: January 2022 until December 2024
- <https://opengpt-x.de>
- Funder: German Federal Ministry for Economic Affairs and Climate Action (BMWK)



Federal Ministry
for Economic Affairs
and Climate Action



OpenGPT-X

Consortium



OpenGPT-X: Developing a Gaia-X node for large AI language models and innovative language application services

Consortium

- Coordinator: **Fraunhofer IAIS/IIS**
- Industry partners: **IONOS, ControlExpert**
- SMEs, startups: **Aleph alpha, Alexander Thamm GmbH**
- Public broadcaster: **WDR**
- Research centres: **Fraunhofer, DFKI, Forschungszentrum Jülich, TU-Dresden**
- Networking: **KI Bundesverband**
- Associated partners (BMW, eco-Verband, Eclipse Foundation, Aalto University, ...)

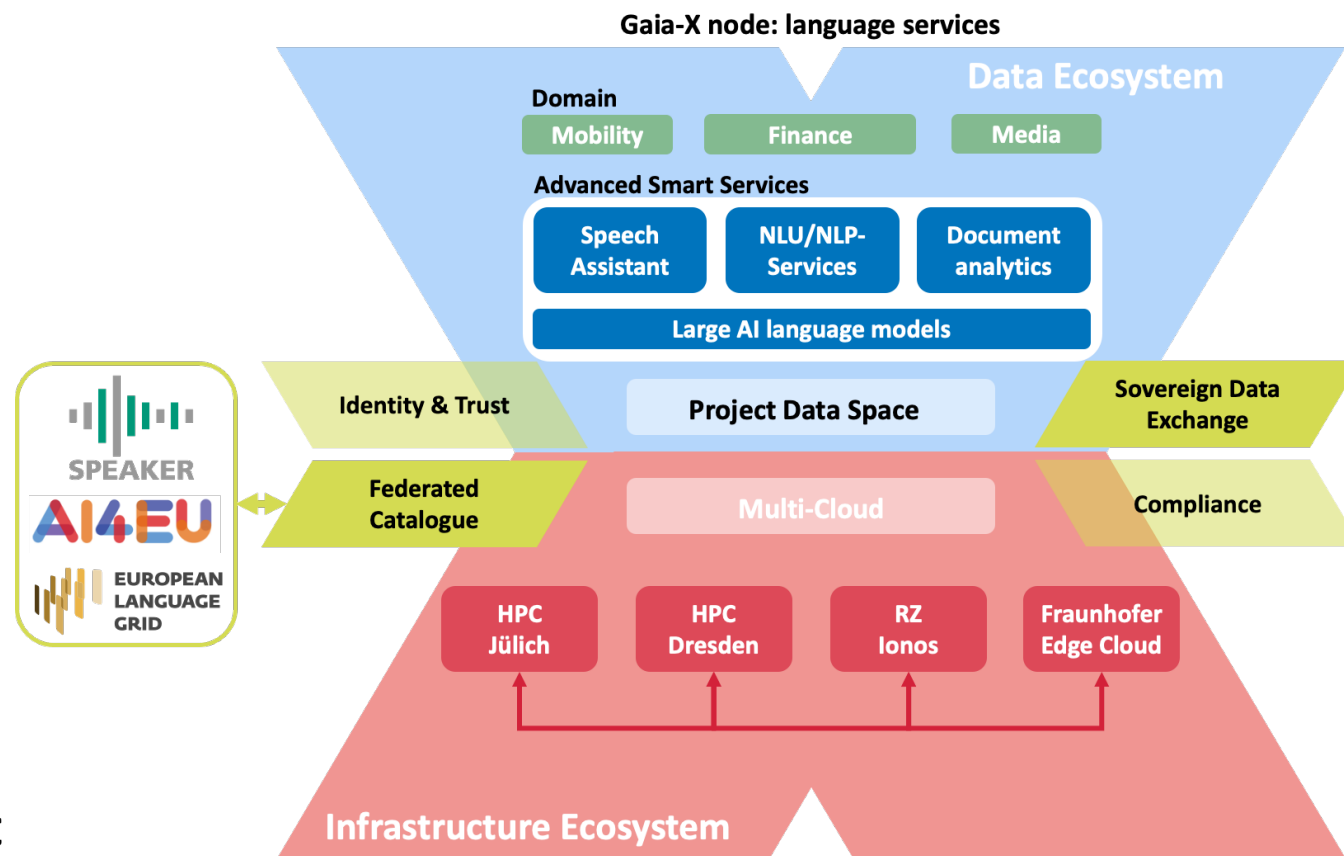


Lighthouse Project OpenGPT-X



Summary

- Development and deployment of **open-source LLMs** and their application as **innovative smart services** in selected **Gaia-X domains** to support **digital sovereignty**
- Build sustainable data and compute infrastructure: **production line for LLMs**
- Development of **innovative AI-based smart language services** for German industry
- **Use cases** developed in OpenGPT-X will be published in the **Gaia-X use case gallery**
- **OpenGPT-X LLM** will be published as **OSS**
- **Total budget: 19 mio. € – Funding: 15 mio. €**

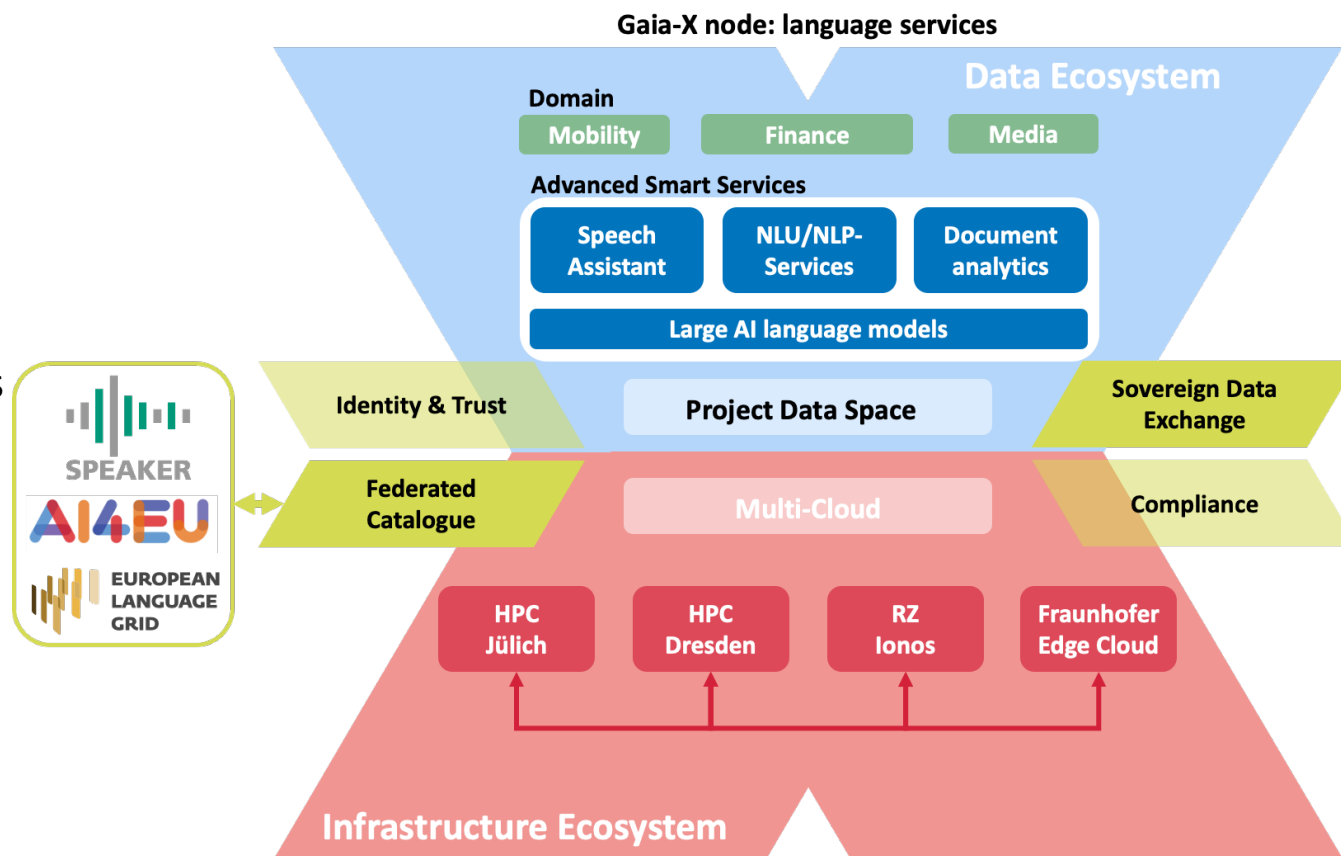


Lighthouse Project OpenGPT-X



Selected Work Packages

- **WP1:** Development of an Gaia-X-enabled, highly scalable GPU-based infrastructure
- **WP2:** Development of an Gaia-X-compatible data infrastructure and MLOps workflows
- **WP3:** Development of large language models
- **WP4:** Development and deployment of innovative AI-based language services
- **WP5:** Development of a Gaia-X node with federated catalogues – interoperability with other AI and LT platforms
- *WP6-WP9 not shown here*



WP1: Gaia-X-enabled GPU infrastructure



- Provide access to HPC resources
 - ZIH Dresden, Ionos, Jülich Supercomputing Centre
- Connection to and interoperability with Gaia-X
- Optimisations on platform level
- Benchmarking and optimisation at hardware level (GPUs)
- Market monitoring and evaluation of GPU and accelerator vendors
- Goal: Create well-performing hardware infrastructure best to train LLMs.



Intermediate Results

Compute time project proposal devised and accepted (approx. four months after project start)

Project partners have **access** to 3rd fastest **supercomputer** in Europe: JUWELS Booster at Forschungszentrum Jülich
➔ First training runs!

Deployment of special hardware at Ionos.

IONOS

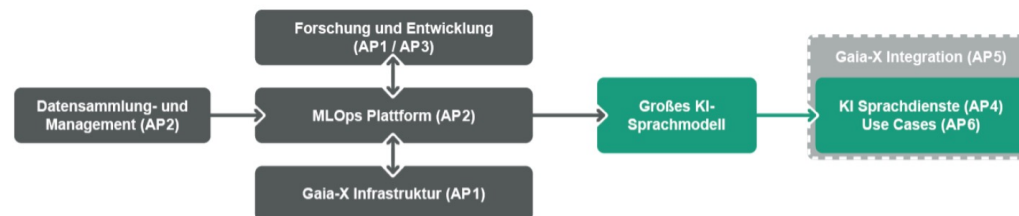
JÜLICH
Forschungszentrum

ZIH
Center for Information Services &
High Performance Computing

WP2: Gaia-X-compatible data infrastructure



- WP2 full title: Development of an Gaia-X-compatible data infrastructure and MLOps workflows
 - Establish an HPC-capable MLOps platform and data management solution in Gaia-X
 - Assembly line architecture for data and infrastructure
 - Includes data collection, data curation and data management
- Challenges:
 - Development of an ML & data infrastructure from scratch in the wider context of an emerging infrastructure and interfaces (Gaia-X)
 - Higher engineering complexity due to duplication of environments
 - Collection of very large corpora of texts and corresponding aligned knowledge graphs
 - Formulating quality standards (language-specific or language-independent)



WP3: Development of large language models



- Multilingual language models: German and English
- Adapt pre-trained language models to new languages
- Knowledge-driven language models
 - Integration of knowledge graphs – how to improve the factual correctness of LLMs?
 - Novel model architectures
- Quantification of bias
- Software architecture
 - Extend the OpenGPT-X system architecture
 - Evaluate existing code bases for reuse: BigScience, OPT, etc.
 - Extend LM Evaluation Harness to incorporate German benchmark datasets

WP3: Development of large language models



- Multilingual language models
 - Train multilingual models from scratch (English, German and a few other EU languages)
 - Evaluation of downstream tasks in German
 - German GPT2-XL – adapted pre-trained language model to English using WECHSEL
 - Word embeddings initialised with WECHSEL, all other weights taken from English gpt2-xl.
 - Training with BigScience's DeepSpeed-Megatron-LM code base for approx. 3 days on 16xA100 GPUs using DFKI's HPC cluster
 - Training dataset: German subset of OSCAR
 - Preliminary evaluation on held-out subset of OSCAR:

Model (size)	PPL
gpt2-xl-wechsel-german (1.5B – ours)	14.5
gpt2-wechsel-german-ds-meg (117M – ours)	26.4
gpt2-wechsel-german (117M)	26.8
gpt2 (retrained from scratch) (117M)	27.6



Model (1.5B parameters) is available on Huggingface (MIT license):

[malteos/gpt2-xl-wechsel-german](https://huggingface.co/malteos/gpt2-xl-wechsel-german)

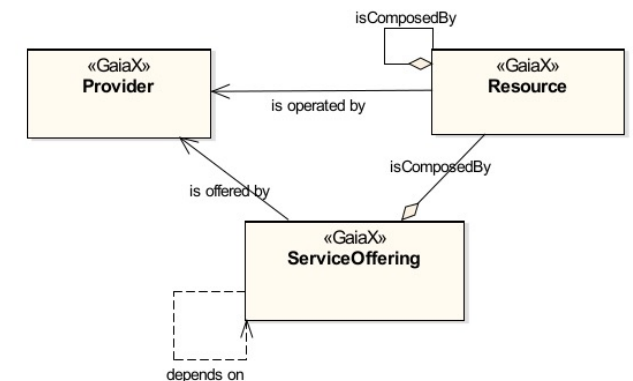
WP4: Innovative AI-based language services



- Goal: Gaia-X-enabled language services based on LLMs
 - Gaia-X self descriptions: Language services can be used within Gaia-X
 - Generative dialog system: Using LLMs for chit-chat and to verbalise factual information
 - Question answering: Verbalise answers in a more natural way
 - Include (use case-specific) knowledge in LLM so that QA can be done entirely in the model
 - Speech Recognition: Improve accuracy by rescoring model output with LLMs and using LLMs in the initial decoding pass of the speech utterance
 - Document analysis: Use LLMs to improve understanding of visually rich documents and integrate LLMs in multi-modal context for layout-aware analysis

WP5: Gaia-X node with LT platform interoperability open**GPT-X**

- Full title: Development of a Gaia-X node with interoperability to established AI/LT platforms
- Goal is to federate LT platforms via Gaia-X
 - OpenGPT-X itself
 - ELG – European Language Grid
 - SPEAKER – Speech Assistant Platform “made in Germany”
 - AI4EU – European AI on Demand Platform
- Requires integrating the (*emerging*) Gaia-X Federation Services (GXFS)
 - Provide the Gaia-X catalogue
 - Create self-descriptions (SD) for Language Technology resources
 - Export all Language Technology resources as Gaia-X entities
- Realised with GXFS prototype
 - Based on (*emerging*) GXFS reference implementation

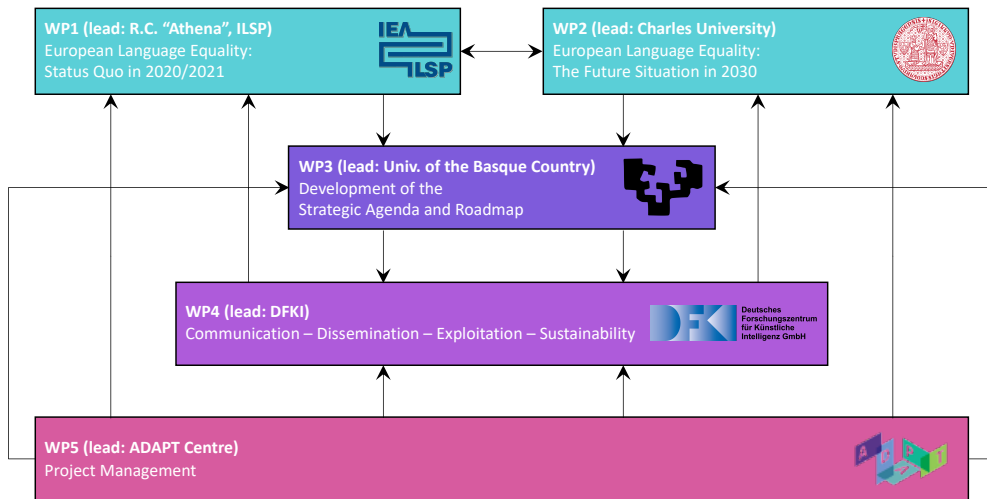


Outline

1. European Language Grid (ELG)
2. OpenGPT-X
- 3. European Language Equality (ELE)**
4. Summary and Conclusions



European Language Equality (ELE)



Consortium: 52 partners from all over Europe

Coordinator: ADAPT Centre (Dublin City University)

Co-Coordinator: DFKI

Objective: *development of a strategic research, innovation and implementation agenda to achieve digital language equality in Europe by 2030*

Runtime: 18 months – ELE and ELG are both finishing up in June 2022

Started on 1 January 2021

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

ELE 2 to run from July 2022 until June 2023

EUROPEAN LANGUAGE TECHNOLOGY

EUROPEAN
LANGUAGE
EQUITY

EUROPEAN
LANGUAGE
GRID

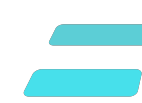


Consortium with 52 partners: 5 core partners, 9 networks and associations, 9 companies, 29 research organisations.

META-FORUM 2022 – June 8/9, Brussels, Belgium

Digital Language Equality Metric

- Digital Language Equality Metric
 - Provide theoretical basis to achieve DLE in Europe by 2030
 - Key ingredient of the evidence-based SRIA and roadmap
 - Evidence: the European Language Grid catalogue
 - Enable comparisons among Europe's languages
 - Identify gaps to help prioritize future interventions
- Modular and flexible design
- Well-defined quantifiers, measures and indicators
- Compatibility and interoperability with ELG metadata schema, i.e., the DLE definition is fully aligned with ELG
- Intended to guide and prioritize future LT development and LR creation, collection, curation



**EUROPEAN
LANGUAGE
EQUALITY**

D1.3

**Digital Language Equality
(full specification)**

Authors	Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, Andy Way
Dissemination level	Public
Date	28-02-2022

Technological Factors & Contextual Factors – Examples

Technological Factors

Example: Corpora/Datasets

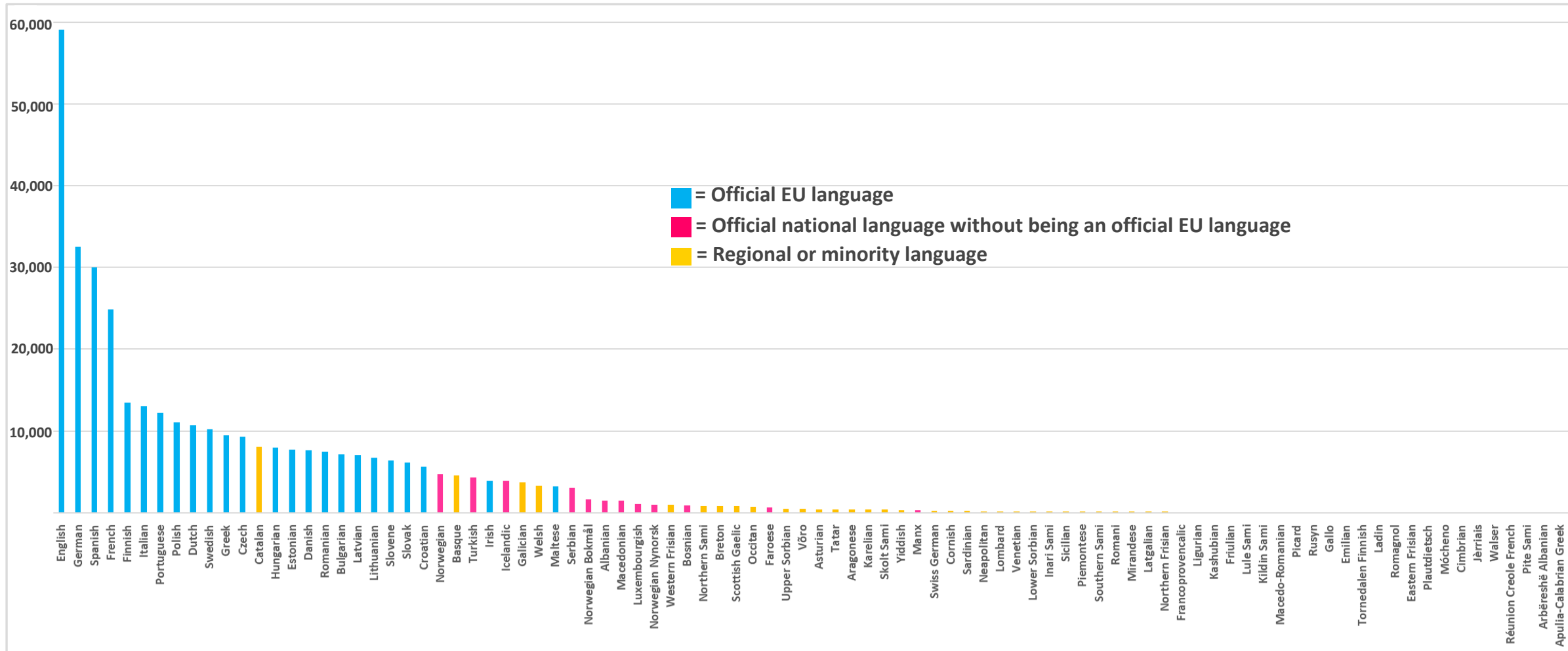
- Language(s)
- Domain(s)
- License
- Type of access
- Corpus subclass
- Media type(s) of parts
- Multilinguality type
- Corpus size

Contextual Factors

Example: Economy

- Size of economy of respective country, countries, region
- Size of LT/NLP market in country, countries, region
- Size of language service and translation/interpreting market in the respective country, countries or region
- Percentage of the IT/ICT sector relative to the economy
- Investment instruments targeting AI/LT/NLP start-ups
- Regional or national LT/NLP/LSP etc. market
- Average socio-economic status of members of the language community

DLE Metric: Technological Scores



Language Reports

- In ELE we wrote 35 language reports
- Authors: approx. 100 colleagues from all over Europe
- Updates of the META-NET White Papers (2012)

D1.4	Report on Basque	UPV/EHU	February 2022	26 pages
D1.5	Report on Bulgarian	IBL	February 2022	27 pages
D1.6	Report on Catalan	BSC	February 2022	24 pages
D1.7	Report on Croatian	FFZG	February 2022	30 pages
D1.8	Report on Czech	CUNI	February 2022	23 pages
D1.9	Report on Danish	UCPH	February 2022	26 pages
D1.10	Report on Dutch	INT	February 2022	23 pages
D1.11	Report on English	USFD	February 2022	21 pages
D1.12	Report on Estonian	UTART	February 2022	20 pages
D1.13	Report on Finnish	UHEL	February 2022	24 pages
D1.14	Report on French	CNRS	February 2022	42 pages
D1.15	Report on Galician	UVIGO	February 2022	20 pages
D1.16	Report on German	DFKI	February 2022	25 pages
D1.17	Report on Greek	ILSP	February 2022	30 pages
D1.18	Report on Hungarian	NYTK	February 2022	26 pages
D1.19	Report on Icelandic	SAM	February 2022	22 pages
D1.20	Report on Irish	DCU	February 2022	30 pages
D1.21	Report on Italian	FBK	February 2022	24 pages
D1.22	Report on Latvian	IMCS	February 2022	29 pages
D1.23	Report on Lithuanian	LKI	February 2022	24 pages

<https://european-language-equality.eu/deliverables/>

Language Reports

Cross-language comparison

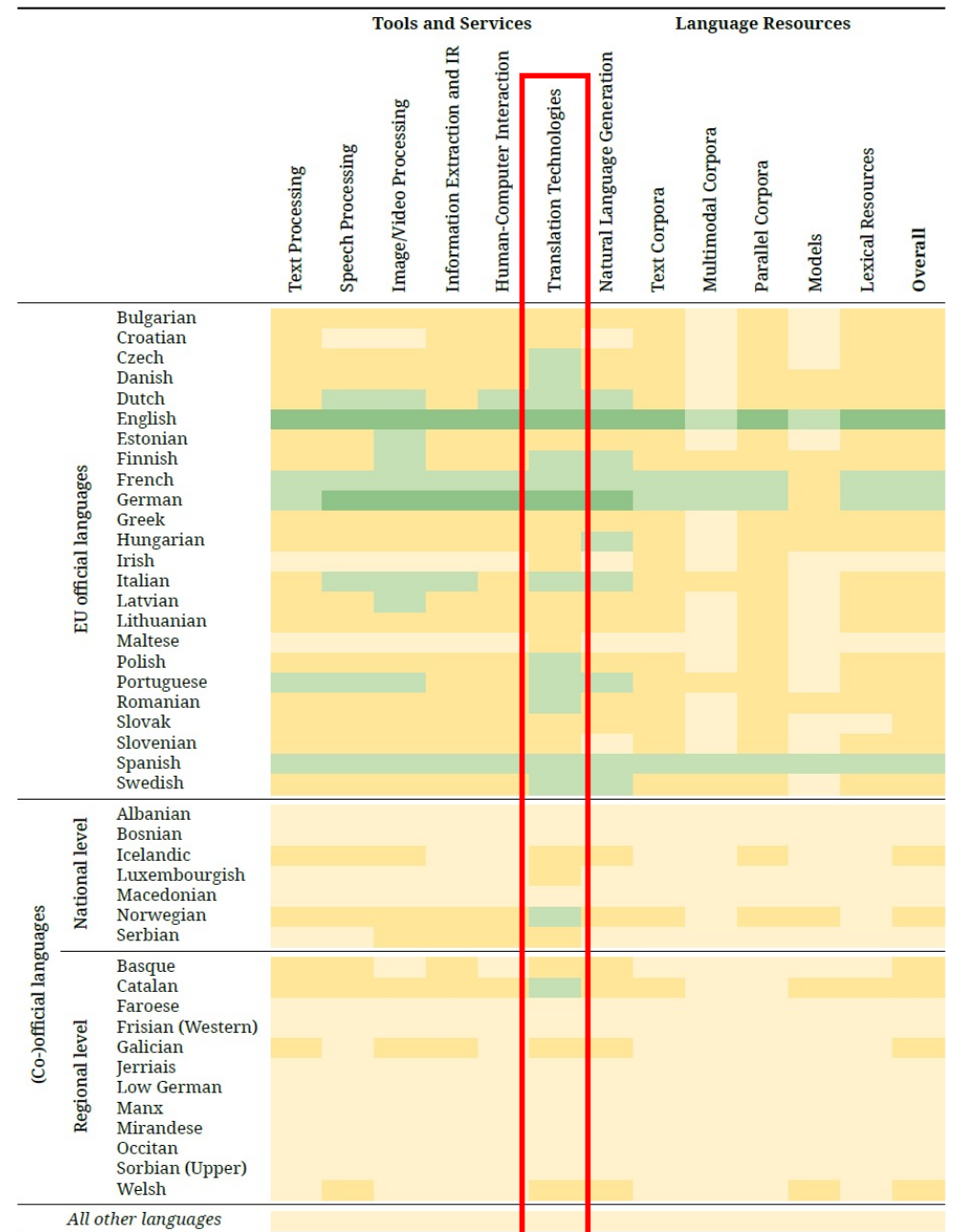
- Based on number of LRTs catalogued in ELG
- Comparison across languages as per
 - Tools/services broadly categorised into a number of core LT application areas
 - Resources that can be used as training or evaluation data (indication of the potential for LT development) with regard to a small number of basic types
- Four bands:
 - Good support (green)
 - Moderate (light green)
 - Fragmentary (yellow)
 - Weak or no support (light yellow)

		Tools and Services							Language Resources					
		Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	Overall
EU official languages	Bulgarian													
	Croatian													
	Czech													
	Danish													
	Dutch													
	English													
	Estonian													
	Finnish													
	French													
	German													
	Greek													
	Hungarian													
	Irish													
	Italian													
	Latvian													
	Lithuanian													
	Maltese													
	Polish													
	Portuguese													
Romanian														
Slovak														
Slovenian														
Spanish														
Swedish														
National level	Albanian													
	Bosnian													
	Icelandic													
	Luxembourgish													
	Macedonian													
	Norwegian													
	Serbian													
Regional level	Basque													
	Catalan													
	Faroese													
	Frisian (Western)													
	Galician													
	Jerriais													
	Low German													
	Manx													
	Mirandese													
	Occitan													
	Sorbian (Upper)													
	Welsh													
All other languages														

Language Reports

Cross-language comparison

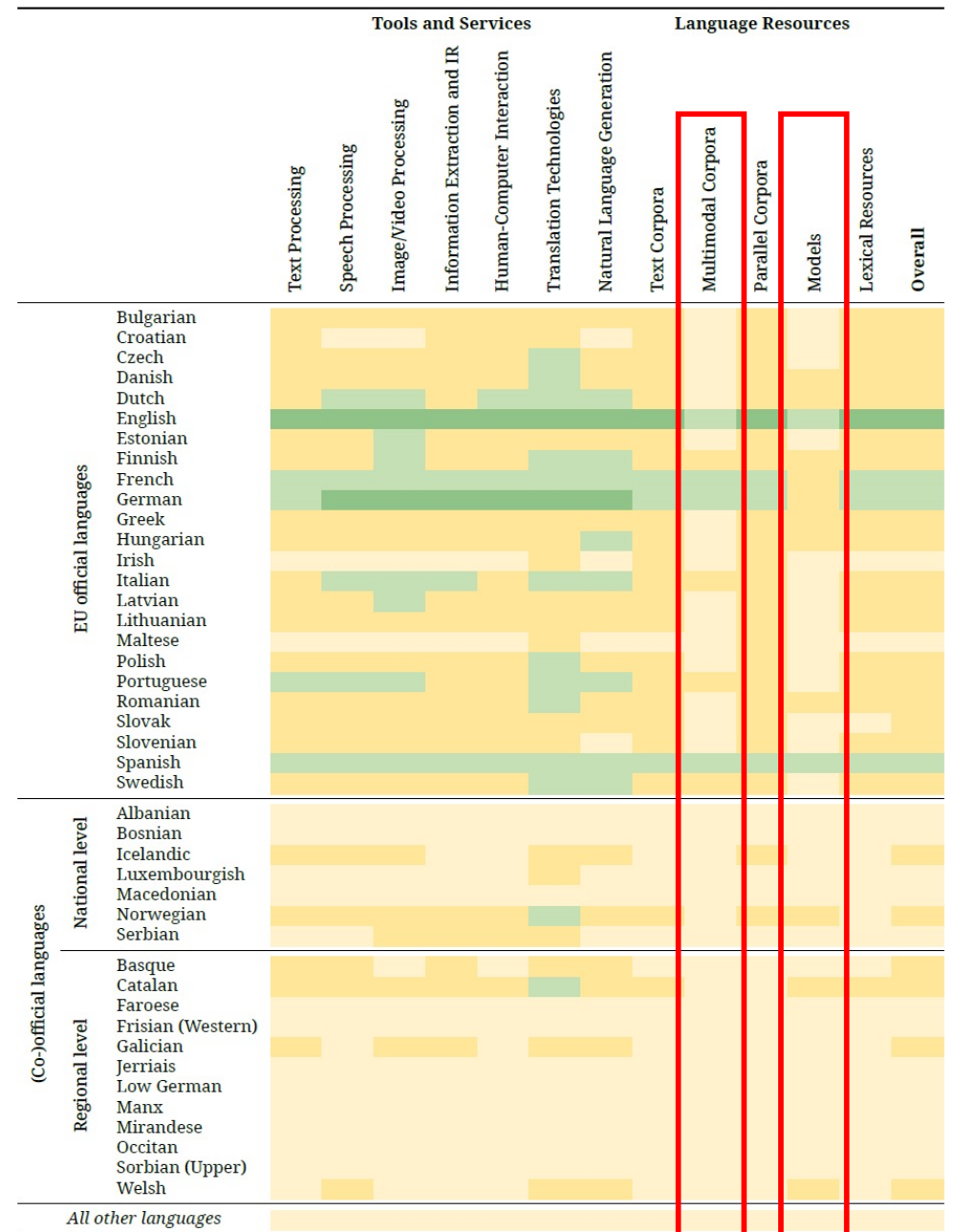
- **Translation technologies:** many languages are at least moderately supported



Language Reports

Cross-language comparison

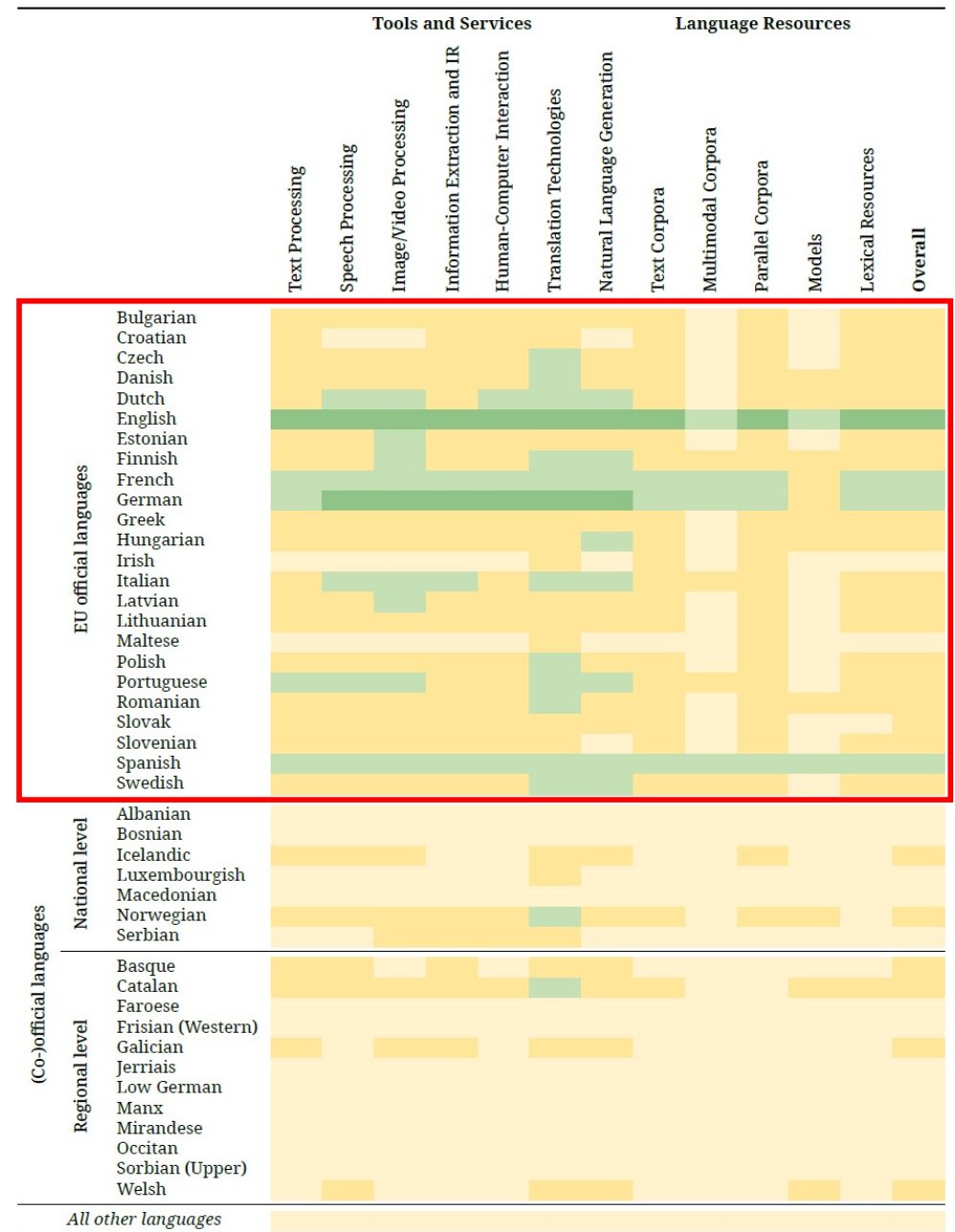
- **Multimodal corpora and language models:** many languages are weakly supported or not supported at all



Language Reports

Cross-language comparison

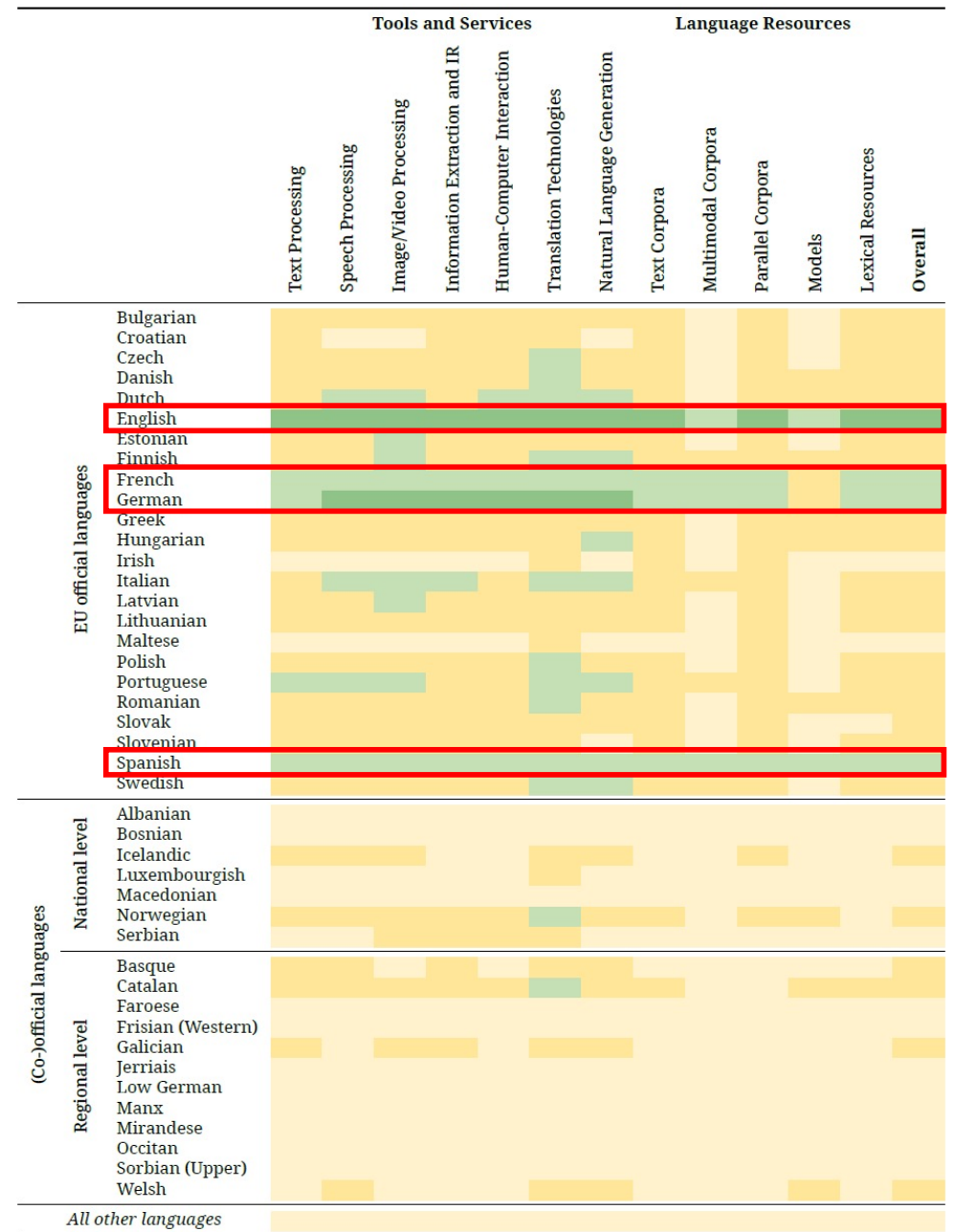
- **EU official languages** better supported than other European languages



Language Reports

Cross-language comparison

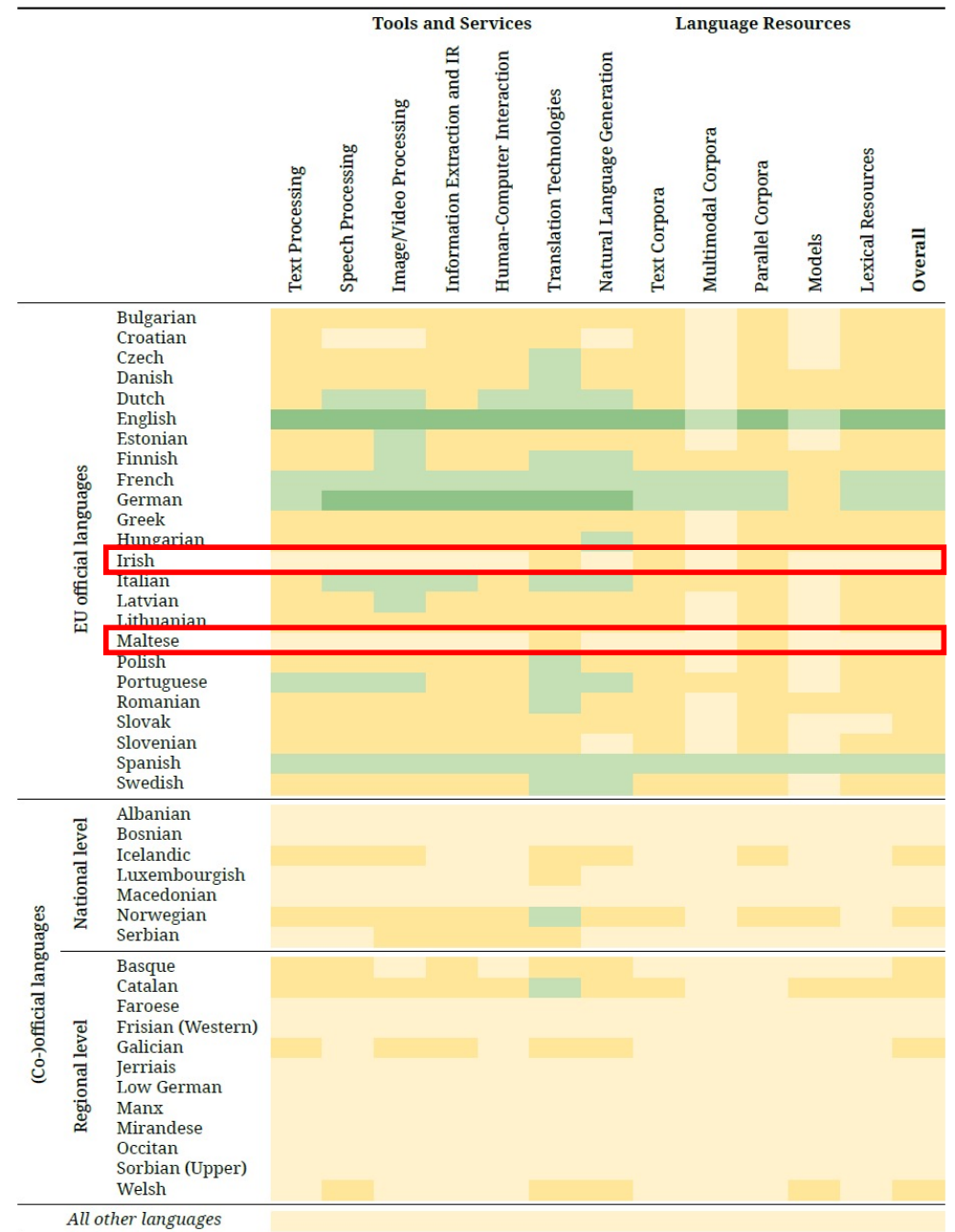
- **Best supported:** English, followed by Spanish, German and French



Language Reports

Cross-language comparison

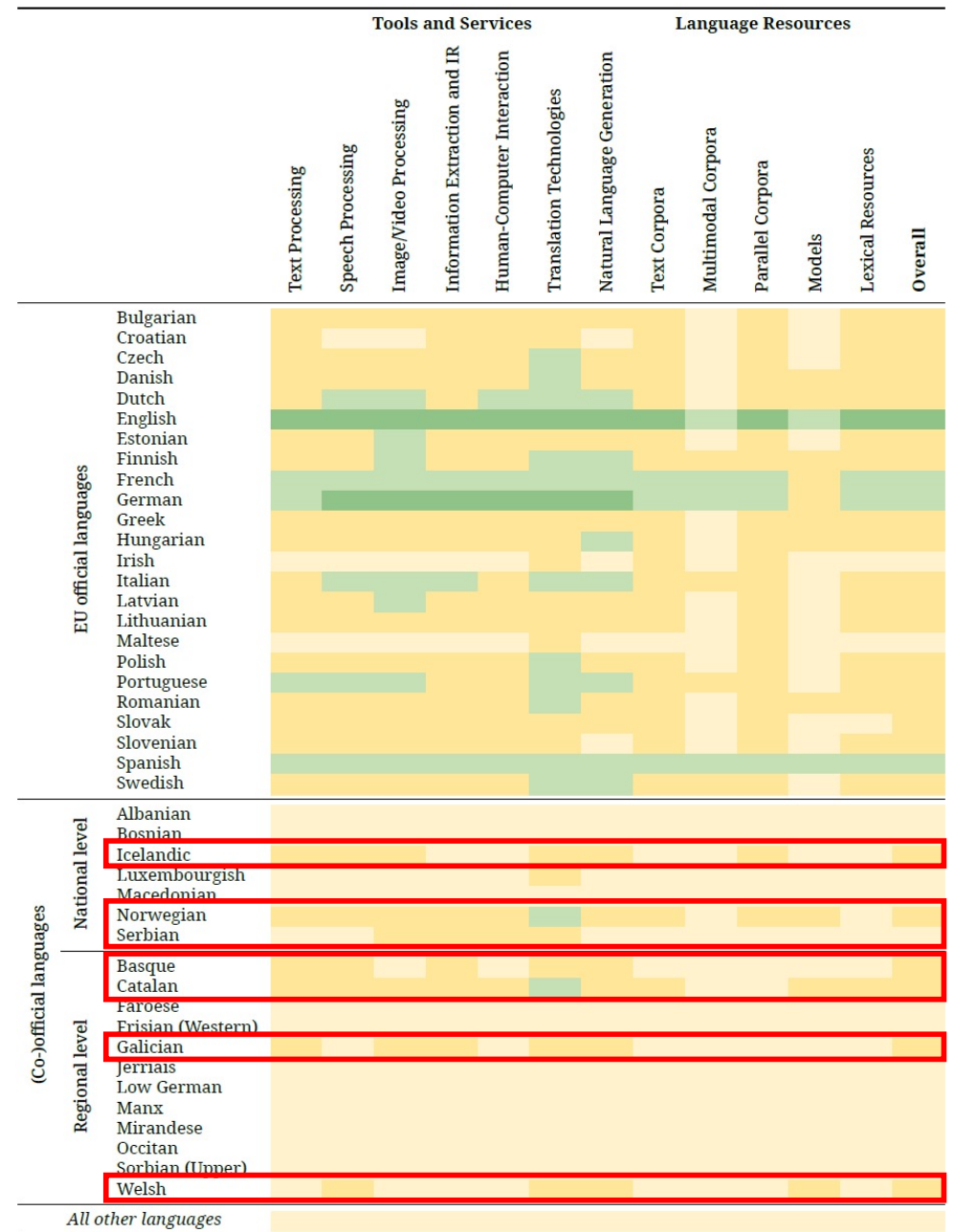
- **Least supported** among the official EU languages:
Irish and Maltese



Language Reports

Cross-language comparison

- Best supported among languages with official status at the national or regional level (but not EU): Norwegian, Icelandic, Serbian, Basque, Catalan, Galician, Welsh



No.	Deliverable name	Short name	Date
D1.1	Digital Language Equality – preliminary definition	DCU	3
D3.1	Report on existing strategic documents and projects in LT/AI	EHU	3
D2.1	Specification of the consultation process including templates, surveys, events etc.	CUNI	4
D1.2	Report on the state of the art in Language Technology and Language-centric AI	EHU	9
D1.3	Digital Language Equality – full specification of the concept	DCU	13
D1.4–D1.35	Report on Basque ... Report Welsh	EHU	14
D2.2–D2.12	Report from CLAIRE ... Report from Wikipedia	ULEID	14
D2.13–D2.16	Technology deep dives MT, Speech, Text Analytics, Data	TILDE	14
D2.17	Report on all external consultations and surveys	CUNI	15
D3.2	Strategic agenda including roadmap – initial version	DFKI	15
D1.36	Database and dashboard with the empirical data collected in D1.4-D1.35 (and others)	ILSP	16
D2.18	Report on the state of Language Technology in 2030	CUNI	16
D3.3	Report on the final round of feedback collection	EHU	17
D3.4	Strategic agenda including roadmap – final version	DFKI	18
D4.5	Strategic agenda and roadmap (print version, online version)	DFKI	18
D4.6	ELE book publication	DFKI	18

Deliverables (condensed to those approx. 55 reports that are relevant for the SRIIA development)

SRIA: Outline

Strategic Research, Innovation and Implementation Agenda and a Roadmap for Achieving full Digital Language Equality in Europe by 2030

1. Multilingual Europe and Digital Technologies
2. Trends and Mega-Trends in Digital Technologies
3. Language Technology and Language-Centric Artificial Intelligence
4. Language Technology and Digital Language Equality in 2022
5. Language Technology and Digital Language Equality in 2030: The ELE Technology Vision and Priority Research Themes
6. A Shared European Programme for Language Technology and Digital Language Equality in Europe by 2030: Recommendations
7. Roadmap towards Digital Language Equality in Europe by 2030

SRIA: Outline

Strategic Research, Innovation and Implementation Agenda and a Roadmap for Achieving full Digital Language Equality in Europe by 2030

1. Multilingual Europe and Digital Technologies
2. Trends and Mega-Trends in Digital Technologies
3. Language Technology and Language-Centric Artificial Intelligence
4. Language Technology and Digital Language Equality in 2022
5. Language Technology and Digital Language Equality in 2030: The ELE Technology Vision and Priority Research Themes
6. **A Shared European Programme for Language Technology and Digital Language Equality in Europe by 2030: Recommendations**
7. Roadmap towards Digital Language Equality in Europe by 2030



Main ELE recommendations

SRIA: Outline

Strategic Research, Innovation and Implementation Agenda and a Roadmap for Achieving full Digital Language Equality in Europe by 2030

6. A Shared European Programme for Language Technology and Digital Language Equality in Europe by 2030: Recommendations

1. Overview and Main Concept
2. Research Recommendations
3. Technology and Data Recommendations
4. Infrastructure Recommendations
5. Policy Recommendations
6. Governance Model

Main language-independent recommendations, extracted from the various ELE deliverables and feedback rounds (more than 2000+ pages of ELE reports).

The whole community contributed to the ELE deliverables, especially through the various ELE surveys amongst the members of the included networks but also through 60+ interviews with experts from the field. We estimate that 500-750 colleagues have contributed to the process.

SRIA: Section 6.2 – Research Recommendations (*draft*)

- Refocus and massively strengthen European LT/NLP research through a large-scale initiative as a shared, collaborative programme between EU and participating countries (*Deep NLU by 2030*).
- Create large open access language models for all European languages: *datasets*, tools, models that include *symbolic knowledge*, *discourse features* and other advanced capabilities.
- Combine long-term encyclopedic knowledge with short-term, learned knowledge representations.
- More interdisciplinary research: enable better modeling of multimodal environments, research how modalities can enrich one another and how training and test sets can be constructed.
- Deep learning and neural approaches: focus on trustworthy, interoperable, explainable LT/NLP/AI.
- Combine interactive LT (conversational AI) with text, knowledge and multimedia technologies for a new generation of applications.

SRIA: Section 6.3 – Technology and Data Recommendations (*draft*)

- Reduce the technology and resource gap between Europe’s languages.
- **Address the *huge* problem of a lack of available data** (relevant for almost *all* European languages).
 - More focus upon **systematic language data collection** (text, dialog, multimodal). Exploit automatic data generation (synthetic data), crowd-sourcing, translation of data.
 - Unleash the power of public sector data, data from broadcasters, social media, publishers etc.
 - Develop methods to overcome the unequal availability of data, by focusing on, e.g., annotation transfer, multilingual models preserving quality, few-shot or zero-shot learning.
- Focus upon **open ecosystems, open source, open access, open standards, interoperability**; foster international **standardisation** of European approaches in LT, NLP, AI, e.g., APIs etc.
- More systematic research in **data bias** including dimensions of bias and how to tackle bias.
- Focus upon **green LT**, i.e., small compute and carbon footprint (e.g., model compression).

SRIA: Section 6.4 – Infrastructure Recommendations (*draft*)

- **Long-term support of infrastructure initiatives – allows implementation of interoperability.**
 - LT infrastructures & data spaces that support research and development activities (such as ELG).
- **Strengthen and reinforce investment into ELG** as the primary European LT platform; enable and foster the sharing of LT resources, services, datasets, models and code between all stakeholders.
- Infrastructures are essential for maximising the exploitation potential of publicly funded research results. **Sufficient operational capacity needs to be ensured, especially for language models.**
- **Infrastructures should provide interoperability** by following standards for data interoperability as well as services and innovative metadata and **data management tools for the whole data life cycle.**
- **Ensure access to HPC and suitable compute infrastructure** (current HPC centres come with too much administrative overhead and do not meet the needs of ML and LT/NLP researchers).

Outline

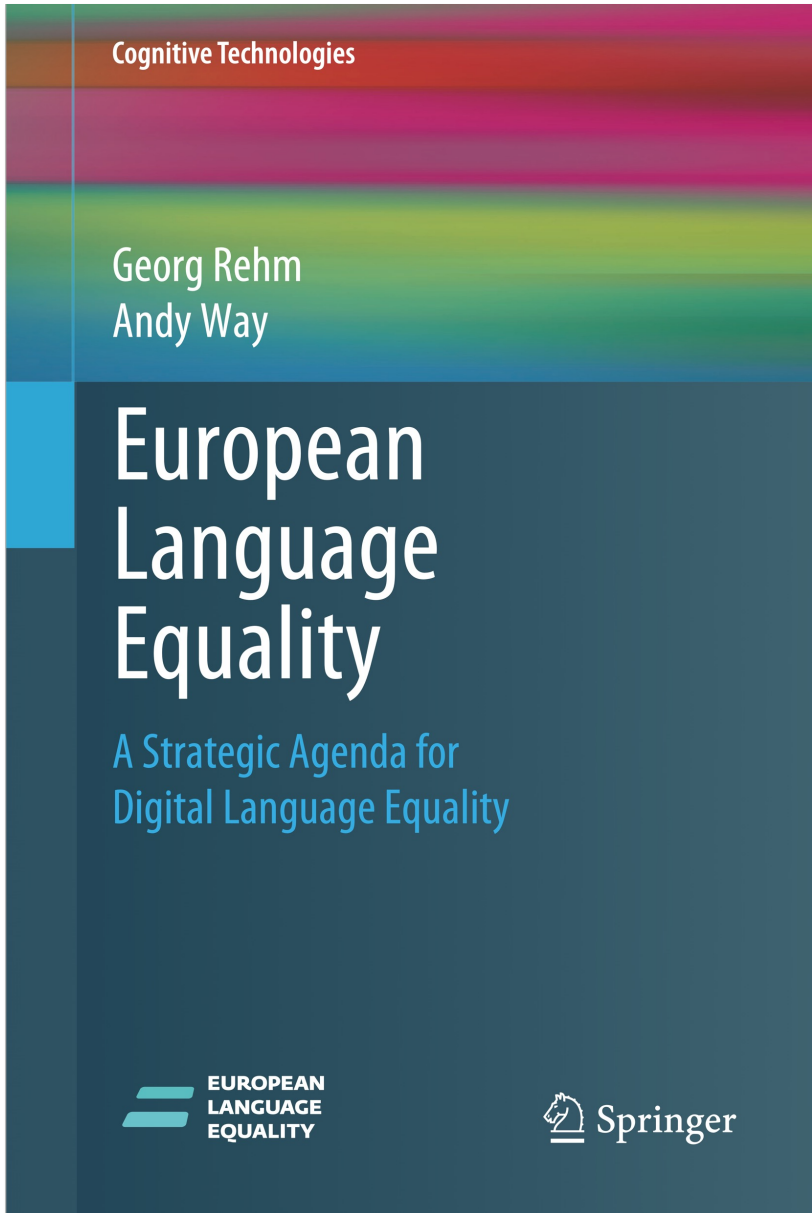
1. European Language Grid (ELG)
2. OpenGPT-X
3. European Language Equality (ELE)
4. **Summary and Conclusions**

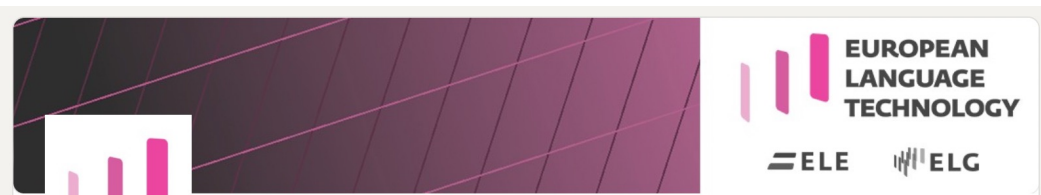


Conclusions and Next Steps



- Global NLP/LT market size by 2025 is enormous: we want Europe to be a key player.
- With ELG we (*finally!*) have a primary platform and marketplace for Language Technology in Europe.
- ELG is a long-term initiative: we will establish a legal entity for sustainability (second half of 2022).
- OpenGPT-X will, among others, integrate the ELG platform into the emerging Gaia-X ecosystem.
- ELE is currently finishing the SRIA and starting ELE 2 in July 2022. *Towards a long-term funding programme!*
- We need much more data, much more easily accessible compute and much more European coordination.
- We need joint principles and scientific goals behind which the European NLP/LT community can assemble.
- Extremely strong demand for LLMs in *all* countries, for *all* European languages.
- ELG has a very strong demand for additional compute (GPUs) for hosting LLMs and offering inference.
- There are various LLM projects (Hungary, Spain, Germany etc.) but no pan-European coordination.
- We need more activity when it comes to unlocking language data – ELE 2 and OpenGPT-X will make a push.
- OpenGPT-X can act as a blueprint for similar LLM-developing projects in other European countries.



European Language Technology
 Fostering the European Language Technology community towards digital language equality in Europe by 2030!
 Research Services · Brussels · 763 followers

[Following](#) [Learn more](#) [More](#)

[Home](#) [About](#) [Posts](#) [Jobs](#) [People](#) [Videos](#)

About

This is the combined channel of the EU projects European Language Equality (ELE) and European Language Grid (ELG) – together working towards a joint network of language technology for Europe’s languages and digital language equality by 2030. Follow us for project updates, events and news from the European artifici... [see more](#)

[See all details](#)

<https://www.linkedin.com/company/european-language-technology>

<https://twitter.com/EuroLangTech>

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

Subscribe to our newsletter

More than 4000 subscribers already!



META-FORUM 2022




European Language Technology
 266 Tweets

European Language Technology
 @EuroLangTech

The channel of the EU projects European Language Equality and European Language Grid, fostering the LT community towards digital language equality by 2030!

📍 Europe european-language-technology.eu 📅 Joined June 2021

1,039 Following 629 Followers

[Edit profile](#)



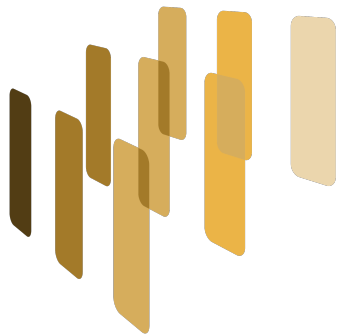
Welcome to European Language Technology!
 | | | ELT

[Follow ELT on Twitter](#)
[Follow ELT on LinkedIn](#)

European Language Technology is the combined communication channel for the sister projects European Language Grid (ELG) and European Language Equality (ELE) – funded by the European Commission.

For further information about either project – their goals, consortium partners, and contact details etc. – please click below:





Thank you!



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

The European Language Equality project has received funding from the European Union under grant agreement № LC-01641480 – 101018166 (ELE),

Prof. Dr. Georg Rehm (DFKI) – Coordinator ELG, Co-Coordinator ELE

14-06-2022 Large language models: pre-training with a twist

<http://www.european-language-grid.eu> – <https://european-language-equality.eu> – <https://opengpt-x.de>

With many thanks to all colleagues in the ELG, ELE and OpenGPT-X teams and consortia!