

European Language Grid – META-FORUM 2022 Kalina Bontcheva – European Language Grid: LT Services and Resources

ELG Release 3



LT Services on the European Language Grid

- ELG Release 3 (June 2022) has close to 2,000 ELG-compatible services (1959)
 - Majority are language dependent
 - 220 are language independent
- APIs for various types of services, including ASR, IE, MT, TTS, OCR
- All 24 official EU languages covered and many more (24 other EU; 69 other languages)
- 1349 services for the official EU languages
 - 765 MT; 472 TA; rest other services
- 198 services for other languages in EU
 - 98 MT; 88 TA; rest other services
- 412 services for other languages
 - 228 TA; 86 MT; rest other services





Dit is een voorbeeld in het Duits

EUROPEAN LANGUAGE GRID

LT Services: From the Pilots

- Many services have been contributed by the ELG-funded pilot projects
- Services include:
 - Over 100 MT language pairs from OPUS-MT
 - Clinical NER in 5 languages from E3C
 - Terminology extraction service from Text2TCS
 - Italian classification services from EVALITA
 - Multilingual WSD, SRL and AMR annotation services (around 100 languages)
 - Sign language explanations







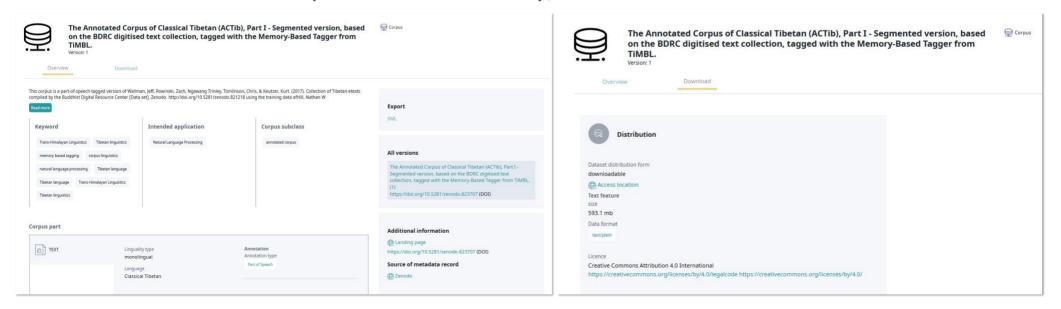
LT Services: from elsewhere

- Attend Session 3 to see details of selected LT services provided by others:
 - Machine Translation:
 - NTEU Pangeanic
 - OPUS-MT University of Helsinki
 - Text-to-speech:
 - Elhuyar Basque ASR/TTS Elhuyar
 - Automatic speech recognition:
 - Welsh ASR Service Bangor University
 - Microservices at your Service Lingsoft
 - A collection of information extraction services, basic linguistic pre-processing, MT



Language Resources: metadata description and harvesting

- ELG has imported metadata from many other repositories:
 - ELRA Catalogue, ELRC-SHARE, ELRA-SHARE-LRs, LINDAT/CLARIAH-CZ, CLARIN Poland repository, CLARIN Slovenia repository, META-SHARE-DFKI, META-SHARE-ELDA and META-SHARE-ILSP, Hugging Face, Quantum Stat, Zenodo
 - Some metadata can be imported automatically, some need manual correction







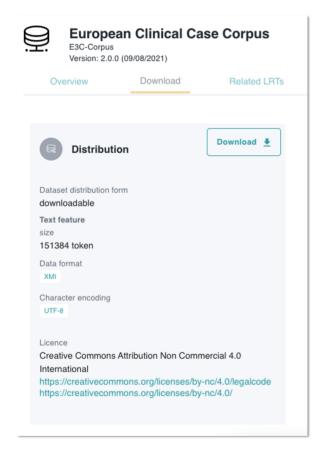
Statistics on ingested LRs that are publicly visible (June 2022)

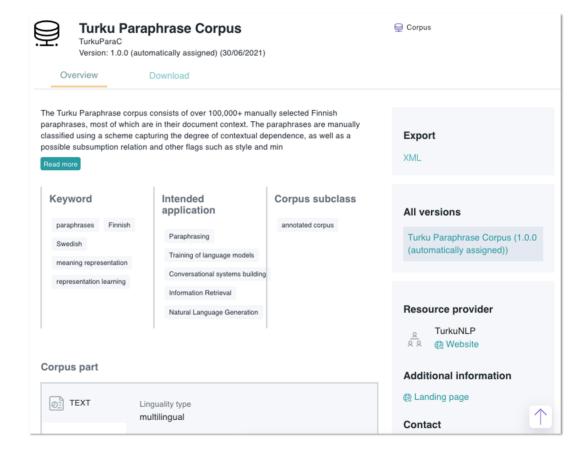
Repository	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
ELRA	635	545	_	1180
ELRC-SHARE	1249	50	_	1299
META-SHARE	52	12	7	71
ELRA-SHARE-LRs	105	37	2	144
LINDAT/CLARIAH-CZ	290	89		400
CLARIN.SI	140	78		218
CLARIN.PL	225	26	31	282
Quantum Stat	255	6	_	261
Zenodo	342	134	37	513
Tudatalib	1	_	_	1
HuggingFace	385	_	_	385
ELG / ELE	2580	1277	367	4224
TOTAL	6259	2254	444	8978



Language Resources: hosted datasets

• Data resources can also be hosted by ELG for direct download, e.g., from pilot projects

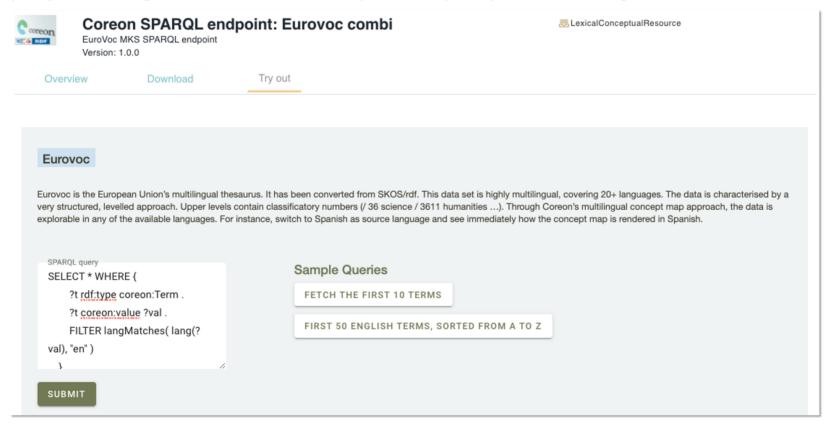






Language Resources: other LR types

- Besides downloadable LRs, ELG also supports LRs accessed via a query interface
- Coreon pilot project integrated SPARQL endpoints queryable through ELG







Language Resources: current state

- Some large and important repositories (Zenodo, HuggingFace) are particularly complex, with sparse or variable quality metadata
 - Harvesting procedures have been implemented whenever possible to optimise ingestion (Zenodo)
- Critical issues being addressed include:
 - Duplication the same resource in multiple repositories
 - Re-use datasets that claim to be distinct but have data in common, e.g., WMT or RumourEval shared tasks
 - Updates new datasets added to already imported repositories
- Legal issues
 - ELG is a complex platform with heterogeneous data flows
 - Need for monitoring for GDPR compliance and a specific data management plan
 - ELG services and datasets use over 130 different licences between them
 - The "Conditions of use" metadata field has been associated to each identified licence (to improve the search functionality)
 - All licences have been analysed and where possible homogenised



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, which will operate and maintain the technology platform for the whole LT community as a joint marketplace.
- Contribute to **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: validation of ELG products; attach more data repositories; include ELG in relevant infrastructures (e.g., NFDI, GAIA-X); establish ELG legal entity.

European Language Grid: An Overview

Georg Rehm¹, Maria Berger¹, Ela Elsholz¹, Stefanie Hegele¹, Florian Kintzel¹, Katrin Marheinecke¹, Stelios Pineridis², Miltos Deligiannis², Dimitrios Galanis², Katerina Gkirtzou², Penny Labronoulou², Kalina Bontcheya³ David Janes Jan Roberts Jan Hajič Jana Hamrlová Lukáš Kačena Khalid Choukri Victoria Arranz Andreis Vasilievs⁶, Orians Anyari⁶, Andis Lagzdinš⁶, Jūlija Melnika⁶, Gerhard Backfried⁷, Erinc Dikici⁷, Miroslay Janosik', Katia Prinz', Christoph Prinz', Severin Stampler', Dorothea Thomas-Anjola', José Manue Gómez Pérez8, Andres Garcia Silva8, Cristian Berrío8, Ulrich Germann9, Steve Renals9, Ondrej Klejch9

1 DFKI GmbH, Germany • 2 ILSP/Athena RC, Greece • 3 University of Sheffield, UK • 4 Charles University. Czech Republic • 5 Evaluations and Language Resources Distribution Agency (ELDA), France • 6 Tilde, Latvia • ⁷ SAIL LABS Technology GmbH, Austria • ⁸ Expert System Iberia SL, Spain • ⁹ University of Edinburgh, UK

Corresponding author: Georg Rehm - georg.rehm@dfki.de

are world this fragm in an easy tools and s IT comm other reso jobs and o Competer

multiling cation in large pla ness is a mains ar holding | guages a digital la rnai. 201 as a unif 2018: Eu The proj addresse the prima commun easy-to-i

Keyword

With 24 Market c Language LT indus outperfor why there veloped t

cluding 1 https:

Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid

Penny Labronoulou¹, Katerina Gkirtzou¹, Maria Gayriilidou¹, Miltos Deligiannis¹, Dimitrios Galanis¹, Stelios Piperidis¹, Georg Rehm², Maria Berger², Valérie Mapelli³, Mickaël Rigault³, Victoria Arranz³, Khalid Choukri³, Gerhard Backfried⁴, José Manuel Gómez Pérez⁵, Andres Garcia Silva⁵

¹ILSP/Athena RC, Greece, ²DFKI GmbH, Germany, ³Evaluations and Language Resources Distribution Agency (ELDA), France, ⁴SAIL LABS Technology GmbH, Austria, ⁵Expert System Iberia SL, Spain {penny, gkirtzou, maria, mdel, galanisd, spip}@athenarc.gr, {georg.rehm, maria.berger}@dfki.de {mapelli, mickael, arranz, choukri}@elda.org, Gerhard.Backfried@sail-labs.com, {agarcia, jmgomez}@expertsystem.com

The current scientific and technological landscape is characterised by the increasing availability of data resources and processing tools and services. In this setting, metadata have emerged as a key factor facilitating management, sharing and usage of such digital assets. In this paper we present ELG-SHARE a rich metadata schema catering for the description of Language Resources and Technologies (processing and generation services and tools, models, corpora, term lists, etc.), as well as related entities (e.g., organizations, projects, supporting documents, etc.). The schema powers the European Language Grid platform that aims to be the primary hub and marketplace for industry-relevant Language Technology in Europe. ELG-SHARE has been based on various metadata schemas, vocabularies, and ontologies, as well as related recommendations and guidelines.

Keywords: metadata, language technology, language technology services, language resources

1. Introduction

The rise of data-driven approaches that use Machine Learning (ML), and especially the breakthroughs in the Deep Learning field, has put data into a central place in all scientific and technological areas. Natural Language Processing (NLP) being no exception. Datasets and NLP tools and services are made available through various repositories (institutional, disciplinary, general purpose, etc.), which makes it hard to find the appropriate resources for one's purposes. Even if they are brought together in one catalogue, such as the European Open Science Cloud1 or the Google dataset search service2 the difficulty of spotting the right resources and services among thousands still remains. Metadata plays an instrumental role in solving this puzzle, as it becomes the intermediary between consumers (humans and machines)

In addition, in the European Union, with the 24 official and many additional languages, multilingualism, cross-lingual and cross-cultural communication in Europe as well as an inclusive Digital Single Market3 can only be enabled and firmly established through Language Technologies (LT). The boosting of the LT domain is thus of utmost importance. To this end, the European LT industry needs to be strengthened promote its products and services integrate them into applications, and collaborate with academia into advancing research and innovation, and bringing research outcomes to a mature level of entering the market. The European Language Grid (ELG) project4 aims to drive forward the European LT sector by creating a platform and establishing it as the primary hub and marketplace for the LT community. The ELG is developed to be a scalable

cloud platform, providing in an easy-to-integrate way, access to hundreds of commercial and non-commercial LTs for all European languages, including running tools and services as well as data resources. Discovery of and access to these resources can only be achieved through an appropriate metadata schema. We present here the ELG-SHARE schema which is used for the description of LT-related resources shared through the ELG platform and its contribu-

2. Objectives

The ELG project (Rehm et al., 2020a) aims to foster European LT by addressing the fragmentation that hinders its development; see indicatively (Rehm and Hegele, 2018; Rehm et al., 2016). To this end, it builds a platform dedicated to the distribution and deployment of Language Resources and Technologies (LRT), aspiring to establish it as the primary platform and marketplace for industry-relevant LT in Europe. The promotion of LT stakeholders and activities and growth of their visibility and outreach is also one of its goals. Together with complementary material in the portal (e.g., training material, information on events, job offerings, etc.), ELG offers a comprehensive picture of the European LT sector.

The ELG platform5 will offer access to hundreds of commercial and non-commercial LTs and ancillary data LRs for all European languages and more: these include processing and generation services, tools, applications for written and spoken language, corpora, lexicons, ontologies, term lists models etc. All resources are accessed through their descriptions in the ELG catalogue LRT providers can describe, upload, and integrate their assets in ELG, and LRT

The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe

Georg Rehm¹, Katrin Marheinecke¹, Stefanie Hegele¹, Stelios Piperidis², Kalina Bontcheva³, Jan Hajič⁴, Khalid Choukri⁵, Andrejs Vasiljevs⁶, Gerhard Backfried⁷, Christoph Prinz⁷, José Manuel Gómez Pérez⁸, Luc Meertens⁹ Paul Lukowicz¹, Josef van Genabith¹, Andrea Lösch¹, Philipp Slusallek¹, Morten Irgens¹⁰, Patrick Gatellier¹ Josephim Köhler¹² Laure Le Bars¹³ Dimitra Anastasiou¹⁴ Albina Auksoriūtė¹⁵ Núria Bel¹⁶ António Branco¹⁷ Gerhard Budin⁸. Walter Daelemans¹⁹. Koenraad De Smedt²⁰. Radovan Garabik²¹. Maria Gavriilidou². Daema Gromann¹⁸, Svetla Koeva²², Simon Krek²³, Cvetana Krstev²⁴, Krister Lindén²⁵, Bernardo Magnini²⁶, Jan Odiik² Maciej Ogrodniczuk28, Eiríkur Rögnvaldsson29, Mike Rosner30, Bolette Sandford Pedersen31, Inguna Skadina3 Marko Tadić³³, Dan Tufiş³⁴, Tamás Váradi³⁵, Kadri Vider³⁶, Andy Way³⁷, François Yvon³⁶

1 DFKI GmbH, Germany • 2 ILSP/Athena RC, Greece • 3 University of Sheffield, UK • 4 Charles University, Czech Republic • 5 ELDA, France • 6 Tilde, Latvia • 7 SAIL LABS Technology GmbH, Austria • 8 Expert System Iberia SL, Spain • 9 CrossLang, Belgium • 10 Oslo Metropolitan University, Norway • 11 Thales Group, France •

European Language Grid: A Joint Platform for the European Language Technology Community

Georg Rehm1, Stelios Piperidis2, Kalina Bontcheva3, Jan Haiič4, Victoria Arranz5, Andrejs Vasiljevs6, Gerhard Backfried7, José Manuel Gómez-Pérez8, Ulrich Germann9, Rémi Calizzano¹, Nils Feldhus¹, Stefanie Hegele¹, Florian Kintzel¹, Katrin Marheinecke¹, Julián Moreno-Schneider1, Dimitrios Galanis2, Penny Labropoulou2, Miltos Deligiannis2, Katerina Gkirtzou², Athanasia Kolovou², Dimitris Gkoumas², Leon Voukoutis², Ian Roberts3, Jana Hamrlová4, Dušan Variš4, Lukáš Kačena4, Khalid Choukri5, Valérie Manelli⁵, Mickaël Rigault⁵, Jūlija Melnika⁶, Miroslav Jánošík⁷, Katia Prinz⁷, Andrés García-Silva8, Cristian Berrío8, Ondrej Klejch9, Steve Renals

¹DFKI GmbH, Berlin, Germany; ²R. C. Athena, ILSP, Greece; ³University of Sheffield, UK; 4Charles University, Czechia; 5ELDA, France; 6Tilde, Latvia; 7HENSOLDT Analytics, Austria; 8Expert System, Spain: 9University of Edinburgh, UK

Corresponding author: georg.rehm@dfki.de

Abstract

Europe is a multilingual society in which dozens of languages are spoken. The only option to enable and to benefit from multilingualism is through Language Technologies (LT), i. e., Natural Language Processing and Speech Technologies. We describe the European Language Grid (ELG), which is targeted to evolve into the primary platform and marketplace for LT in Europe by providing one umbrella platform for the European LT landscape, including research and industry enabling all stakeholders to upload, share and distribute their services products and resources. At the end of our EU project, which will establish a legal entity in 2022, the ELG will provide access to apnrox 1300 services for all European languages as well as thousands of data sets

same tii few yea present

Keywo

and int

a large

mately

tilingu

guages

ing sev

ticale

etrating

Europe is a multilingual society with 24 EU Member State languages and dozens of additional languages including regional and minority languages and languages spoken by immigrants, trade part- is also meant to support digital language equalners and tourists. The only option to enable and ity in Europe (STOA, 2017; European Parliament, to benefit from multilingualism is through Lan- 2018), i.e., to create a situation in which all language Technologies (LT) including Natural Lan- guages are supported through technologies equally guage Processing (NLP) and Speech Technologies well. The current imbalance is characterised by a (Rehm, 2017). While the European LT landscape stark predominance of LTs for English, while alis world class, it is also massively fragmented most all other languages are only marginally sup-(Vasilievs et al., 2019; Rehm et al., 2020d).

guage Grid (ELG) cloud platform.1 This scal- Rehm et al., 2014, 2016a; ELRC, 2019). able system is targeted to evolve into the primary

platform for LT in Europe. It will provide one umbrella platform for all LTs developed by the European LT landscape, including research and industry, addressing a gap that has been repeatedly raised by the European LT community for many years (Rehm and Uszkoreit 2013: Rehm et al., 2016b; STOA, 2017; Rehm, 2017; Rehm and Hegele, 2018; European Parliament, 2018). ELG is meant to be a virtual home and marketplace for all products, services and organisations active in the LT space in Europe (Rehm et al., 2020a). The platform can be used by all stakeholders to showcase, share and distribute their products, services. tools and resources. At the end of the EU project ELG (2019-2022), which will establish a legal entity in early 2022, the platform will provide access to approx. 1300 commercial and non-commercial tools and services for all European languages, as well as thousands of language resources (LRs). ELG will enable the European LT community to deposit and upload their technologies and data sets and to deploy them through the grid. The ELG ported and, thus, in danger of digital language ex-We describe Release 2 of the European Lan- tinction (Rehm and Uszkoreit, 2012; Kornai, 2013;

Section 2 gives an overview of the ELG plathttps://www.european-language-grid.eu. We provide a screencast demo video at https://youtu.be/LDQadki/z/M. related work. Section 4 concludes the peace related work.

Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability

Georg Rehm¹, Dimitrios Galanis², Penny Labropoulou², Stelios Piperidis², Martin Welß³, Ricardo Usbeck³, Joachim Köhler³, Miltos Deligiannis², Katerina Gkirtzou², Johannes Fischer⁴, Christian Chiarcos⁵, Nils Feldhus¹, Julián Moreno-Schneider¹, Florian Kintzel¹, Elena Montiel⁶, Víctor Rodríguez Doncel⁶, John P. McCrae⁷, David Laqua³, Irina Patricia Theile3, Christian Dittmar4, Kalina Bontcheva8, Ian Roberts8, Andreis Vasilievs9, Andis Lagzdins9

1 DEKL GmbH. Germany • 2 II SP/Athena RC. Greece • 3 Fraunhofer IAIS. Germany • 4 Fraunhofer IIS. Germany • 5 Goethe University Frankfurt, Germany • 6 Universidad Politécnica de Madrid, Spain • National University of Ireland Galway, Ireland • 8 University of Sheffield, UK • 9 Tilde, Latvia

Corresponding author: Georg Rehm - georg.rehm@dfki.de

Abstract

LREC 2020 Workshop Language Resources and Evaluation Conference 11-16 May 2020

IWLTP 2020 - 1st International Workshop on Language Technology Platforms

PROCEEDINGS

Coming up:

Due ficia socie tinue

ceive nom platf tiona

and Unic

their tors thog platf dom writt alise

- European Language Grid book
- European Language Equality book Plus: ELG Online Documentation

Edited by:

Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajič, Stelios Piperidis, Andrejs Vasiljevs

₩ELG META-FORUM 2022

46

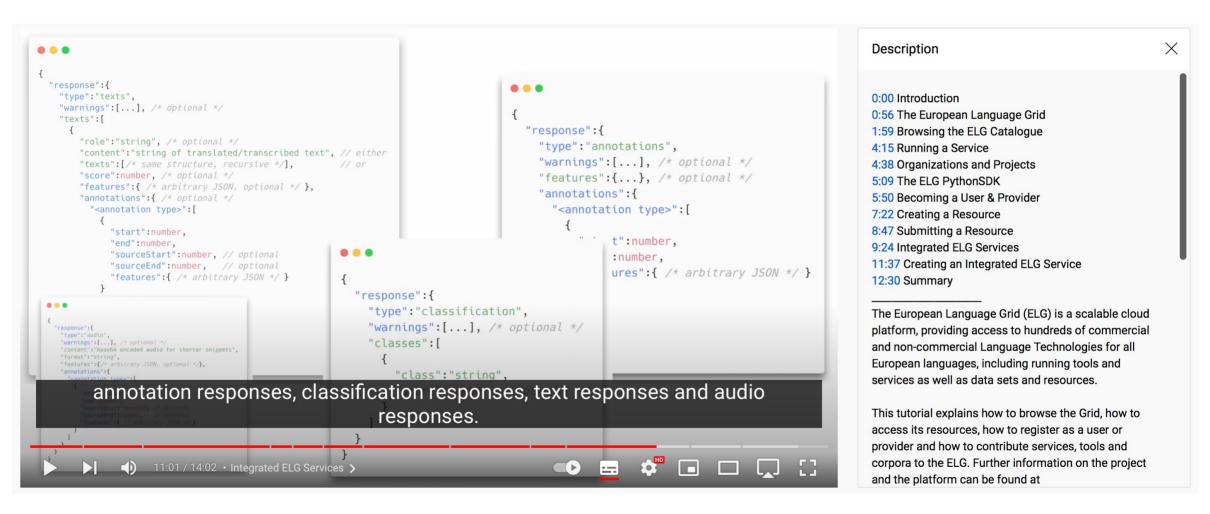
¹https://www.eosc-portal.eu

https://toolbox.google.com/datasetsearc

³https://ec.europa.eu/digital-single-market/en 4https://www.european-language-grid.eu

⁵The ELG platform has just been launched (alpha release) and (official release dates are on April of 2020, 2021 and 2022.

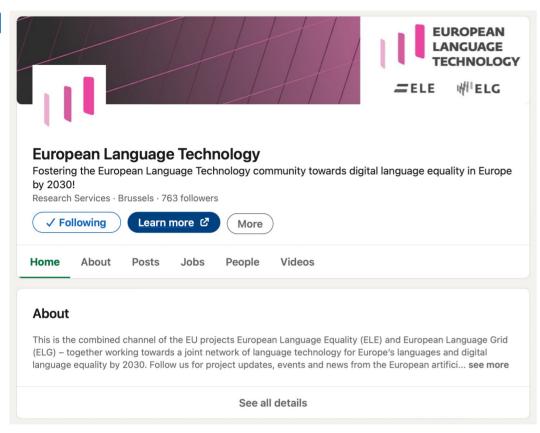
ELG Tutorial Video



https://www.youtube.com/watch?v=29-V2EyMn4E







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!



1.039 Following 629 Followers











META-FORUM 2022

2022	09:00-10:30	Session 1: Opening – European Language Grid		
	10:30-10:45	Coffee Break		
	11:00-11:45	Session 2: ELG Platform		
	11:45-12:30	Session 3: Selected ELG Tools, Services and Resources		
	12:30-13:30	Lunch Break + Video Expo		
June 2	13:30-14:45	Session 4: Language-centric AI Panel		
08	14:45-15:15	Coffee Break		
	15:15-16:15	Session 5: Industry Session		
	16:15-16:45	Session 6: The Future of ELG		
	16:45-17:00	Closing Session		
	17:00-18:00	Reception		



META-FORUM 2022:

Joining the European Language Grid
Together towards Digital Language Equality

Detailed Programme at https://www.european-language-grid.eu/meta-forum-2022

META-FORUM 2022



European Language Grid European Language Equality





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),

Georg Rehm (DFKI), Stelios Piperidis (ILSP, R.C. "Athena"), Kalina Bontcheva (University of Sheffield), georg.rehm@dfki.de, spip@athenarc.gr, k.bontcheva@sheffield.ac.uk

08/09-06-2022 META-FORUM 2022 – Joining the European Language Grid (hybrid conference) http://www.european-language-grid.eu – http://european-language-equality.eu