Chapter 1
European Language Grid: Introduction

Georg Rehm

Abstract Europe is a multilingual society with 24 European Union Member State languages and dozens of additional languages including regional and minority languages as well as languages spoken by immigrants, trade partners and tourists. While languages are an essential part of our cultural heritage, language barriers continue to be unbreachable in many situations. The only option to enable and to benefit from multilingualism is through Language Technologies (LTs) including Natural Language Processing (NLP), Natural Language Understanding (NLU) and Speech Technologies. The commercial European LT landscape is dominated by hundreds of SMEs that develop many different kinds of LTs. While the industrial and also the academic European LT community is world-class, it is also massively fragmented. This chapter is an introduction to the present volume, which describes the European Language Grid (ELG) cloud platform, initiative and EU project. The ELG system is targeted to evolve into the primary platform and marketplace for LT in Europe by providing one umbrella platform for the entire European LT community, including research and industry, enabling all stakeholders to showcase, share and distribute their services, tools, products, datasets and other resources. At the time of writing, the ELG platform provides access to more than 13,000 commercial and non-commercial language resources and technologies covering all official EU languages and many national, co-official, regional and minority languages.

1 Overview and Context

Europe is a multilingual society with 24 EU Member State languages and dozens of additional languages including regional and minority languages as well as languages spoken by immigrants, trade partners and tourists. While languages are an important part of our cultural heritage, language barriers continue to be unbreachable in many situations. The only option to enable and to benefit from multilingualism is through Language Technologies (LTs) including Natural Language Processing (NLP), Nat-

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_1

ural Language Understanding (NLU), and Speech Technologies. The commercial European LT landscape is dominated by hundreds of SMEs and a few larger enterprises (Rehm et al. 2020b). While the European LT community is world class, it is also very fragmented, significantly holding back its impact (Vasiljevs et al. 2019).

This book is the definitive documentation[1] of the EU project *European Language Grid*, which has developed the ELG cloud platform (Figure 1), available online at:

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

The European Language Grid is targeted to evolve into the primary platform for Language Technology in Europe. We provide one umbrella platform for all LTs and LRs developed by the whole European LT landscape, including research and industry, addressing a major gap, i. e., the lack of a common LT platform, that has been repeatedly raised by the whole community for many years (Rehm and Uszkoreit 2013; Rehm et al. 2016; STOA 2018; Rehm and Hegele 2018; European Parliament 2018). The ELG platform is also meant to be a virtual home and marketplace for all products, services and organisations active in this space in Europe, significantly boosting the EU Digital Single Market by helping to make it multilingual. ELG is an initiative *from* the European LT community *for* the European LT community. It provides one platform that can be used by all stakeholders to showcase, share and distribute their products, services, tools, datasets, corpora and other relevant resources. At the time of writing, the ELG platform enables access to more than 13,000 commercial and non-commercial language resources and technologies for all official EU languages and many national, co-official, regional and minority languages.

The European LT community had been demanding a dedicated LT platform for years – the ELG cloud platform fills this gap. The ambition of the ELG project and initiative is to unite a strong and extensive network of European experts and concentate on *commercial* as well as *non-commercial LTs*, both *functional* (analysis, processing and generation for written and spoken language) and *non-functional* (datasets, corpora, lexicons, models etc.). A related goal is to establish the ELG as a marketplace for the fragmented European LT landscape (Vasiljevs et al. 2019; Rehm et al. 2020b) to connect demand and supply, strengthening Europe's position in this field. The ELG platform enables the whole European LT community to upload their services and datasets, to deploy them, connect with, and make use of those resources made available by others (taking into account IPR and licences, as soon as the ELG legal entity is in place, including payment and billing options, especially with regard to commercial services and resources).

ELG is also meant to support *digital language equality* in Europe (STOA 2018; European Parliament 2018), i. e., bringing about a situation in which *all* languages are supported through technologies equally well. Currently, there is still an extreme imbalance, characterised by a stark predominance of LRTs for English, while almost all other languages are only marginally supported (Gaspari et al. 2022; Grützner-Zahn and Rehm 2022). In fact, many of these languages are in danger of digital

[1] The ELG cloud platform is actively being used, i. e., new services, tools and resources are made available on or through ELG on a daily basis. The data, numbers and statistics presented in this book regarding the use of ELG reflect the respective time of writing.

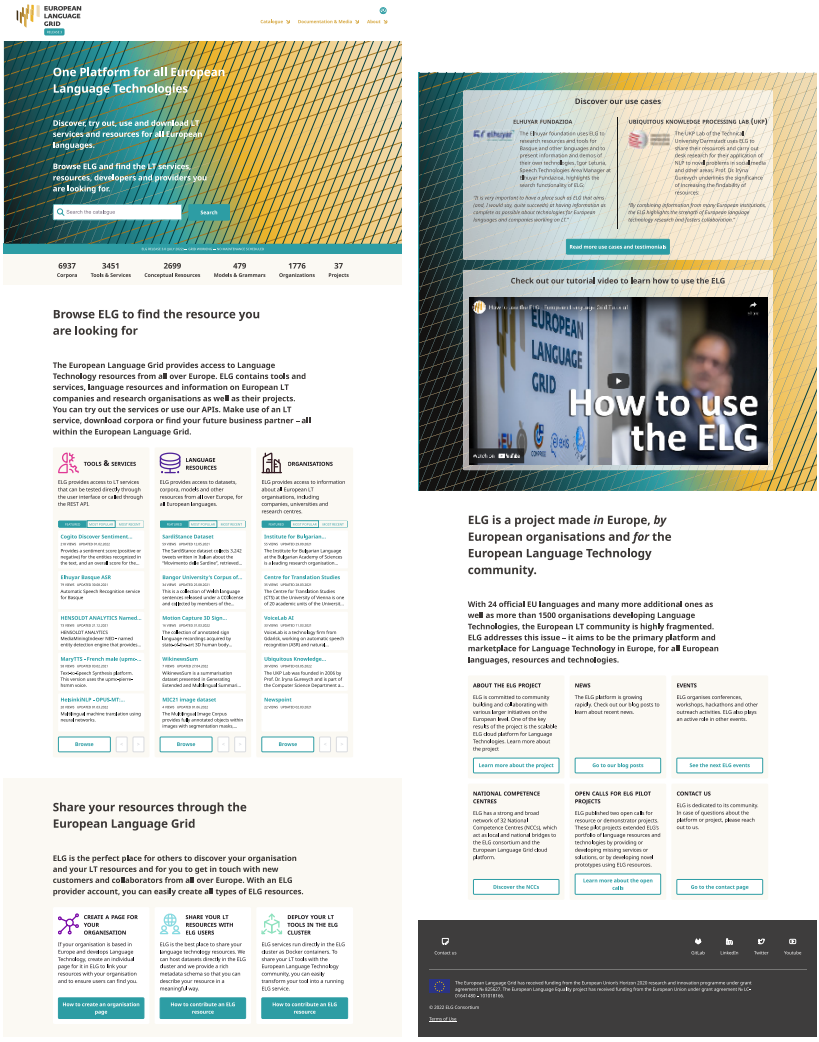

Fig. 1 The European Language Grid cloud platform

language extinction (Rehm and Uszkoreit 2012; Kornai 2013). With an initial consortium of 52 partners, ELG's sister project ELE (European Language Equality; Jan. 2021 – June 2022) and its immediate follow-up project ELE 2 (July 2022 – June 2023) are developing a strategic agenda and roadmap for digital language equality in Europe by 2030 to address this issue by means of a coordinated, pan-European research, development and innovation programme (Rehm and Way 2023).[2]

[2] https://european-language-equality.eu

2 The European Language Grid EU Project

The original proposal for the Innovation Action "European Language Grid" (ELG) was prepared by a consortium of nine partners (Table 1) and submitted on 17 April 2018, responding to the European Commission Horizon 2020 call topic ICT-29-2018 ("A multilingual Next Generation Internet", sub-topic "European Language Grid").[3] The ELG EU project[4] started in January 2019 and finished in June 2022.[5]

1	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Coordinator)	DFKI	Germany
2	Athena Research and Innovation Center in Information, Communication and Knowledge Technologies, Institute for Language and Speech Processing	ILSP	Greece
3	University of Sheffield	USFD	UK
4	Charles University	CUNI	Czech Republic
5	Evaluations and Language Resources Distribution Agency	ELDA	France
6	Tilde SIA	TILDE	Latvia
7	HENSOLDT Analytics GmbH	HENS	Austria
8	Expert System Iberia SL	EXPSYS	Spain
9	University of Edinburgh	UEDIN	UK

Table 1 Consortium of the ELG EU project

The project was structured into three broader *areas*. The *ELG Platform* area (WP 1, WP 2, WP 3) took care of developing the technology platform, which was built with robust, scalable, reliable and widely used open source technologies, enabling it to scale with the growing demand and supply. As an important part of the platform, the ELG catalogue contains metadata records of all resources (including services, datasets etc.), service and application types, languages as well as records of LT companies, research organisations, projects, etc. This is where the first area overlapped with the second, i. e., *ELG Content* (WP 4, WP 5), referring to the actual content of the European Language Grid in terms of processing or generation services, tools, datasets, corpora, models, language resources etc. We distinguished between *functional* content (running services that can be uploaded into and deployed from the ELG cloud platform and integrated into other systems) and *non-functional* content (datasets, corpora, lexicons, etc.). Functional LT services are created by containerising and ingesting them into ELG. One of our key goals was to make this process as easy and efficient as possible for commercial and non-commercial LT providers. These are two of the main classes of users of the third area, i. e., *ELG Community* (WP 6, WP 7), which includes all stakeholders of the ELG. Apart from commercial or academic developers of LT, these stakeholders also include companies, NGOs or

[3] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/ict-29-2018

[4] https://cordis.europa.eu/project/id/825627

[5] The original runtime of 36 months was extended by six months due to the COVID-19 pandemic.

public administrations interested in purchasing or integrating Language Technologies into their own systems and applications. The ELG project collaborated – and still collaborates – with various other EU-supported research and innovation projects as well as with international networks and associations. Furthermore, ELG established a network of 32 National Competence Centres (NCCs) in as many European countries, who acted as national bridges to the project, generating interest in participating in the ELG initiative amongst relevant stakeholders from their own regions. In 2020, ELG published two open calls through which a total of 15 pilot projects were financially supported. These pilot projects extended ELG's catalogue with relevant services or datasets and realised innovative applications based on the ELG platform and available services and resources, demonstrating the usefulness of the platform. Table 2 shows all work packages of the ELG EU project.

Area	Work Package	Lead
ELG Platform	WP 1 Base Infrastructure	DFKI
	WP 2 Language Grid	ILSP
	WP 3 Interactive Interface and Information System	TILDE
ELG Content	WP 4 Services, Tools, Components	USFD
	WP 5 Language Resources, Data Sets and Models	ELDA
ELG Community	WP 6 Piloting the ELG	CUNI
	WP 7 Communication and Competence Centres	DFKI
	WP 8 Project Management and Coordination	DFKI

Table 2 Work packages of the ELG EU project

The ELG project resulted in more than 40 deliverables, the public ones of which are available online.[6] In addition to what had been originally specified in the project plan in early 2018, the project also worked on a number of activities that were not foreseen to be executed in the project proposal or grant agreement. For example, ELG organised the First International Workshop on Language Technology Platforms (IWLTP 2020).[7] Driven by the success of this workshop (Rehm et al. 2020a), a special issue of the *Language Resources and Evaluation* journal focusing on LT Platforms is currently in preparation, scheduled to be published in 2023. Motivated by the very positive feedback we have received from many different stakeholders since the beginning of the project, we decided, in 2020, to compile the present book as the definitive documentation of the project.

[6] https://www.european-language-grid.eu/deliverables

[7] https://www.european-language-grid.eu/iwltp-2020

3 Beyond the ELG EU Project

Throughout the years it has been repeatedly argued that Europe should not outsource its multilingual communication and digital language infrastructure to other continents and markets since the European demands are complex, challenging and above all unique. Instead, Europe should support and make use of its own LT community. One of the obstacles to overcome along the way has been the development of a shared technology and community platform for all European stakeholders. Now that the ELG cloud platform is finally in place, it is able to foster Language Technologies *for Europe* built *in Europe*, tailored to our languages and cultures and to our societal and economical demands, benefitting European citizens, society, innovation and industry. ELG plays the role of a shared, scalable cloud platform for the whole European LT community and it also functions as a joint marketplace and broker for a broad variety of services, products and datasets.

The ELG EU project was successfully completed in June 2022, and Release 3 of the ELG platform is ready to be used. At the time of writing, ELG provides access to more than 13,000 commercial and non-commercial language resources and technologies for all official EU languages and many national, co-official, regional and minority languages. In addition, the ELG project has contributed to validating and extending the platform with 15 pilot projects, building a pan-European community of users and providers, establishing communication and outreach channels and organising a number of large-scale conferences and smaller workshops.

Since the start of the project, we have been collaborating with the European AI on demand platform, especially with the AI4EU project, to ensure compatibility of our approaches in terms of describing resources semantically. Furthering these collaborative efforts will facilitate cross-platform search and discovery enabling ELG resources and other assets to be visible, discoverable and usable by the wider AI community. Considering the EU's plan to deploy the emerging European AI on demand platform, ELG is ready to act as the central language-related AI hub and marketplace providing access to and direct use of several thousands of LT services and datasets.

The ELG legal entity will take over further development and maintenance of ELG in the second half of 2022. At the same time, the ELG platform plays a role in several new funded projects. ELE (Jan. 2021 – June 2022) and ELE 2 (July 2022 – June 2023) have already been mentioned – ELG's sister projects are developing a strategic agenda and roadmap for achieving full digital language equality in Europe by 2030.[8] The ELG platform was and is heavily used in ELE – of special importance is the ELE dashboard, which provides a number of visualisations of the ELG catalogue, enabling various comparisons of the technology support of Europe's languages.[9] The project OpenGPT-X (Jan. 2022 – Dec. 2024), funded by the German Federal Ministry for Economic Affairs and Climate Action, develops large language models that will enable new data-driven business solutions, specifically address-

[8] https://european-language-equality.eu

[9] https://live.european-language-grid.eu/catalogue/dashboard

ing European needs.[10] In this project, many different language resources provided by ELG are used for research and development purposes. In addition, ELG will be further extended so that it complies to the specifications of the emerging Gaia-X[11] infrastructure and ecosystem, eventually integrating ELG into Gaia-X, making available many of the OpenGPT-X results (and *all* ELG resources) through Gaia-X. The project NFDI4DataScience and Artificial Intelligence (Oct. 2021 – Sept. 2026) is part of the initiative *Nationale Forschungsdateninfrastruktur* (German Research Data Infrastructure).[12] In this project, the ELG platform will be integrated into the emerging NFDI[13] infrastructure. A similar goal will be addressed by the upcoming EU project SciLake (Jan. 2023 – Dec. 2025), in which we will establish technical bridges between the ELG platform and the European Open Science Cloud (EOSC).[14] Finally, the upcoming EU project DataBri-X (Oct. 2022 – Sept. 2025) will interlink ELG and the emerging DataBri-X platform.

4 Summary of this Book

This book is structured into four different parts. Parts I, II and III describe the main results of the ELG project, while Part IV focuses on the ELG open calls and the 15 pilot projects. Below we include short summaries of the four parts.

4.1 Part I: ELG Cloud Platform

Part I provides an in-depth description of the *European Language Grid Cloud Platform*. First, Chapter 2 (p. 13 ff.) introduces the architecture and setup of the ELG cloud platform, including fundamental concepts such as the user and provider roles, the semantic metadata scheme and the different types of technologies currently supported by the platform. Afterwards, Chapter 3 (p. 37 ff.) concentrates on using ELG as a *consumer*. For this purpose, the web-based user interface, the public-facing APIs and the ELG Python SDK can be used. The complementary Chapter 4 (p. 67 ff.) examines using ELG as a *provider* of Language Technologies and Language Resources including the corresponding dashboard, service integration and various helper tools. Chapter 5 (p. 95 ff.) goes even deeper and provides a description of the ELG cloud infrastructure, e. g., the Kubernetes cluster, the storage solution etc. Finally, Chapter 6 (p. 107 ff.) examines the relation between ELG and other projects and infrastructures in terms of various technical collaborations (e. g., metadata harvesting).

[10] https://opengpt-x.de

[11] https://gaia-x.eu

[12] https://www.nfdi4datascience.de

[13] https://www.nfdi.de

[14] http://eosc.eu, https://eosc-portal.eu

4.2 Part II: ELG Inventory of Technologies and Resources

Part II focuses on the actual content of the ELG platform, i. e., it examines the *ELG Inventory of Technologies and Resources*. First, Chapter 7 (p. 131 ff.) describes the hundreds of functional Language Technology tools and services available in the ELG platform, covering machine translation, automatic speech recognition, text-to-speech synthesis as well as text analysis tools, among others. These tools and services have been and are being provided by companies as well as academic organisations. Chapter 8 (p. 151 ff.) then takes a look at the diverse set of Language Resources covering datasets, corpora, language models and other types of resources for all European languages. Many of these are hosted in ELG, available for direct download, while for others metadata records are collected from external repositories, enabling discovery through ELG as a one-stop-shop platform for the European LT community. Chapter 9 (p. 171 ff.) concludes Part II and describes the organisations, i. e., companies and research institutions, as well as projects currently represented in ELG. Our vision is for ELG to become the primary platform for Language Technology in Europe and, thus, for all organisations that develop LT to actively maintain their ELG pages, provide language tools and services as well as language resources, linking them to their own ELG pages.

4.3 Part III: ELG Community and Initiative

Part III provides an in-depth look at four different dimensions of the *ELG Community and Initiative*. First, Chapter 10 (p. 189 ff.) describes the main group of stakeholders that the EU project ELG collaborated with including various LT providers, different EU and national research projects as well as several wider initiatives. This chapter also describes the different ELG communication channels including social media. Chapter 11 (p. 205 ff.) focuses on the 32 National Competence Centres (NCCs) that the ELG project set up. The NCCs function as an international network of national networks, they support the overall mission of the ELG project. On a more abstract level, Chapter 12 (p. 219 ff.) provides a glimpse at various aspects and processes that revolve around open innovation and the marketplace concept as one of the main visions we have for the European Language Grid. Finally, Chapter 13 (p. 233 ff.) describes the ELG legal entity – including setup, challenges, products etc. – as the main instrument to sustain the ELG initiative beyond the EU project.

4.4 Part IV: ELG Open Calls and Pilot Projects

Part IV is dedicated to the *ELG Open Calls and Pilot Projects*. A considerable amount of the overall budget of the EU project European Language Grid was set aside to support a number of pilot projects that either make use of the technologies

and resources provided by ELG or that extend the ELG inventory and portfolio by contributing additional technologies or resources. First, Chapter 14 (p. 257 ff.) describes the setup of the ELG open calls including designed and implemented procedures, boards, evaluation criteria etc. The following 15 chapters – Chapter 15 (p. 271 ff.) to Chapter 29 (p. 355 ff.) – report on the 15 pilot projects, selected from more than 200 project proposals in an expert-driven evaluation procedure.

References

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. URL: http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way (2022). "Introducing the Digital Language Equality Metric: Technological Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 1–12. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.1.pdf.

Grützner-Zahn, Annika and Georg Rehm (2022). "Introducing the Digital Language Equality Metric: Contextual Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 13–26. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.2.pdf.

Kornai, Andras (2013). "Digital Language Death". In: *PLoS ONE* 8.10. DOI: 10.1371/journal.pone.0077056. URL: https://doi.org/10.1371/journal.pone.0077056.

Rehm, Georg, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs, eds. (2020a). *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Marseille, France. URL: https://www.aclweb.org/anthology/volumes/2020.iwltp-1/.

Rehm, Georg and Stefanie Hegele (2018). "Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs". In: *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA, pp. 3282–3289. URL: https://aclanthology.org/L18-1519.pdf.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020b). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani,

Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rehm, Georg and Hans Uszkoreit, eds. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg, New York, Dordrecht, London: Springer. URL: http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

Rehm, Georg, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asunción Moreno, Jan Odjik, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Sandford Pedersen, Inguna Skadiņa, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiļjevs, Kadri Vider, and Jolanta Zabarskaite (2016). "The Strategic Impact of META-NET on the Regional, National and International Level". In: *Language Resources and Evaluation* 50.2, pp. 351–374. DOI: 10.1007/s10579-015-9333-4. URL: http://link.springer.com/article/10.1007/s10579-015-9333-4.

Rehm, Georg and Andy Way, eds. (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Forthcoming. Springer.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. URL: https://data.europa.eu/doi/10.2861/136527.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*. DOI: 10.2759/142151. URL: https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.