Chapter 1
European Language Equality: Introduction

Georg Rehm and Andy Way

Abstract This chapter provides an introduction to the EU-funded project *European Language Equality* (ELE). It motivates the project by taking a general look at multilingualism, especially with regard to the political equality of all languages in Europe. Since 2010, several projects and initiatives have developed the notion of utilising sophisticated language technologies to unlock and enable multilingualism technologically. However, despite a landmark resolution that was adopted by the European Parliament in 2018, no significant progress has been made. Together with the whole European LT community, and making use of a concerted community consultation process, the ELE project produced strategic recommendations that specify how to bring about full digital language equality in Europe and reach the scientific goal of Deep Natural Language Understanding by 2030, not only addressing but eventually solving the problem of *digital inequality* of Europe's languages.

1 Overview and Context

In Europe's multilingual setup, all 24 official EU languages are granted equal status by the EU Charter and the Treaty on EU. Furthermore, the EU is home to over 60 regional and minority languages which have been protected and promoted under the European Charter for Regional or Minority Languages (ECRML) treaty since 1992, in addition to various sign languages and the languages of immigrants as well as trade partners. Additionally, the Charter of Fundamental Rights of the EU under Article 21 states that, "[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."

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

Andy Way
Dublin City University, ADAPT Centre, Ireland, andy.way@adaptcentre.ie

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_1

Unfortunately, language barriers still hamper cross-lingual communication and the free flow of knowledge and thought across language communities and continue to be unbreachable in many situations. While multilingualism is one of the key cultural cornerstones of Europe and signifies part of what it means to be and to feel European, no EU policy has been proposed to address the problem of language barriers.

Artificial Intelligence (AI), Natural Language Processing (NLP), Natural Language Understanding (NLU), Language Technologies (LTs), and Speech Technologies (STs) have the potential to enable multilingualism technologically but, as the META-NET White Paper Series *Europe's Languages in the Digital Age* (Rehm and Uszkoreit 2012) found in 2012, our languages suffer from an extreme imbalance in terms of technological support. English is very well supported through technologies, tools, datasets and corpora, for example, but languages such as Maltese, Estonian or Icelandic have hardly any support at all. In fact, the 2012 study assessed *at least 21 European languages to be in danger of digital extinction*. If, as mentioned above, all European languages are supposed to be on an equal footing in general, technologically, they clearly are not (Kornai 2013).

After the findings of the META-NET study and a set of follow-up projects, studies and recommendations (e. g., Rehm and Uszkoreit 2013; STOA 2018), the joint CULT/ITRE report *Language Equality in the Digital Age* (European Parliament 2018) was eventually passed with an overwhelming majority by the European Parliament on 11 September 2018. It concerns the improvement of the institutional framework for LT policies at the EU level, EU research and education policies to improve the future of LTs in Europe, and the extension of the benefits of LTs for both private companies and public bodies. The resolution also recognises that there is an imbalance in terms of technology support of Europe's languages, that there has been a substantial amount of progress in research and technology development and that a large-scale, long-term funding programme should be established to ensure full technology support for all of Europe's languages. The goal is to enable multilingualism technologically since "the EU and its institutions have a duty to enhance, promote and uphold linguistic diversity in Europe" (European Parliament 2018).

While the resolution was a important milestone for the idea of enabling Europe's multilingualism technologically and bringing every language in Europe to the same level of technology support, there has been no concrete follow-up action along the lines laid out in the resolution, i. e., to set up "a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands". In the meantime, however, many highly influential breakthroughs in the area of language-centric AI have been achieved, mostly by large enterprises in the US and Asia, especially approaches and technologies concerning large language models (LLMs such as BERT or ChatGPT).[1]

Due to a lack of action over the last five to seven years, Europe has mostly been playing "second fiddle" in the area of language-centric AI and Language Technolo-

[1] ChatGPT was released in Nov. 2022, https://chat-gpt.org. Most chapters of this book were written by mid-2022, which is why they do not reflect the widespread impact and subsequent recognition of this novel application.

gies. Driven by the "European Strategy for data", the EU is currently concentrating on setting up a number of sectorial data spaces to enable and support the data economy and to boost its digital sovereignty.[2] These, fortunately, also include a dedicated language data space with a focus on stakeholders from industry. *But, simply put, language is much more than data*. In addition to the complex and long-term activity of constructing the aforementioned data spaces, the EU also invests in AI-related actions that include language, albeit with limited budgets. However, much more needs to be done to properly address the challenge of Europe's multilingualism with meaningful and long-lasting solutions.

With a consortium of 52 partners, the EU project *European Language Equality* (ELE; Jan. 2021 – June 2022) and its follow-up project ELE 2 (July 2022 – June 2023) developed, through a large-scale, community-driven process, a *Strategic Research, Innovation and Implementation Agenda for Digital Language Equality in Europe by 2030* to address this major issue by means of a coordinated, pan-European research, development and innovation programme.[3] This book is the definitive documentation of the EU project ELE. It describes the current situation of technology support for Europe's languages and our overall recommendations of what more needs to be done to achieve Digital Language Equality (DLE) in Europe by 2030.

2 The European Language Equality Project

The original proposal for the EU project "European Language Equality" was prepared by a consortium of 52 partners[4] (see Figure 1) and submitted on 29 July 2020, responding to the European Commission call topic PPA-LANGEQ-2020 ("Developing a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030").[5] The ELE project started in January 2021 and finished in June 2022. Immediately after the end of the first ELE project, the one-year ELE 2 project began with a reduced consortium of seven partners, continuing some of the work strands of the first project.

Developing a strategic agenda and roadmap for achieving full DLE in Europe by 2030 involves many stakeholders, which is why the process of preparing the different parts of the strategic agenda and roadmap – the key objective and result of the project – was carried out together with all 52 partners of the consortium and the wider European LT community. We concentrated on two distinct but related aspects: 1. describing the current state of play (as of 2021/2022) of LT support for the languages under investigation; and 2. strategic and technological forecasting, i. e., estimating and envisioning the future situation ca. 2030. Furthermore, we distinguished between two main stakeholder groups: 1. *LT developers* (industry and research) and

[2] https://digital-strategy.ec.europa.eu/en/policies/strategy-data

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

[4] https://european-language-equality.eu/consortium/

[5] https://ec.europa.eu/research/participants/data/ref/other_eu_prog/other/pppa/wp-call/call-fiche_pppa-langeq-2020_en.pdf

Fig. 1 Members of the ELE consortium at META-FORUM 2022 in Brussels (9 June 2022)

2. *LT users and consumers*. Both groups were represented in ELE with several networks, initiatives and associations who produced one report each, highlighting their own individual needs, wishes and demands towards DLE. The project's industry partners produced four in-depth reports compiling the needs, wishes and visions of the European LT industry. We also organised a larger number of surveys (inspired by Rehm and Hegele 2018) and consultations with stakeholders not directly represented in the consortium.

With the development of the strategic agenda, the project followed two complementary goals. 1. The *socio-political goal* was the preparation of a strategic agenda explaining how Europe can bring about full digital language equality by 2030. This objective and the need for a corresponding large-scale, long-term programme have been recognised already by the EU (European Parliament 2018). 2. Additionally, the strategic agenda and the eventual large-scale, long-term funding programme are also meant to pursue a *scientific goal*, i. e., reaching *Deep Natural Language Understanding* by 2030. As briefly mentioned, Europe is currently lagging behind the breakthroughs achieved on other continents, which is why the dedicated large-scale, long-term funding programme we envision can and must achieve both objectives: develop resources and technologies to fully unlock and benefit from multilingualism technologically and also put Europe back into the pole position in the area of LT, NLP and language-centric AI research.

Operationally, the project was structured into five work packages (see Figure 2). In WP1, "European Language Equality: Status Quo in 2020/2021", a definition of the concept of DLE was prepared and the current state-of-the-art in the research area of LT and language-centric AI was documented in a report. The heart of WP1 was the preparation of more than 30 language reports, each documenting one European language and the level of technology support it had as of 2022. While WP1 examined the status quo, WP2, "European Language Equality: The Future Situation in 2030" looked into the future. Operationalised through a complex community consultation

process, we collected and analysed the demands, needs, ideas and wishes of European LT developers (industry and research), European LT users and consumers as well as European citizens. Four technical deep dives took a detailed look at the four main areas of LT (Machine Translation, Speech, Text Analytics and Data). The results of WP1 and WP2 were fed to WP3, "Development of the Strategic Agenda and Roadmap", in which the overall strategic agenda was developed based on the collected findings of WP1 and WP2, including an additional feedback loop with the wider community. WP4, "Communication – Dissemination – Exploitation – Sustainability" organised a number of events, including META-FORUM 2022[6] in Brussels (see Figure 1) and a workshop in the European Parliament.[7] WP4 also set up and managed our social media channels and a newsletter under the umbrella brand "European Language Technology".[8] WP5 took care of managing the large consortium of 52 partners. Figure 3 shows the overall timeline of the project.

Our methodology was, thus, based on a number of stakeholder-specific surveys as well as collaborative document preparation that also involved technology forecasting. Both approaches were complemented through the collection of additional input and feedback through various online channels. The two main stakeholder groups (LT developers and LT users/consumers) differ in one substantial way: while the group of commercial or academic LT developers is, in a certain way, *closed* and well represented through relevant organisations, networks and initiatives in the ELE consortium, the group of LT users is an *open* set of stakeholders that is only partially represented in our consortium. Both stakeholder groups have been addressed with targeted and stakeholder-specific surveys.

The ELE project resulted in around 70 deliverables, of which the public ones are available online.[9] In addition, a number of reports were prepared pro bono by collaborators who supported the goals of the project, including language reports on Bosnian, Serbian, West Frisian, the Nordic minority languages and Europe's sign languages. All reports are available on the ELE website.

3 Beyond the ELE Project

While forecasting the future of the field of LT and language-centric AI is surely an enormous challenge, we can confidently predict that even greater advances will be achieved in all LT research areas and domains in the near future (Rehm et al. 2022). However, despite claims of human parity in many LT tasks, *Deep Natural Language Understanding*, the main scientific goal of the ELE Programme, is still an *open research problem* far from being solved since all current approaches have

[6] https://www.european-language-grid.eu/events/meta-forum-2022

[7] https://www.europarl.europa.eu/stoa/en/events/details/towards-full-digital-language-equality-i/20220711WKS04301

[8] The social media channels and the newsletter were organised in close collaboration with ELE's sister project European Language Grid (ELG, Rehm 2023).

[9] https://www.european-language-equality.eu/deliverables

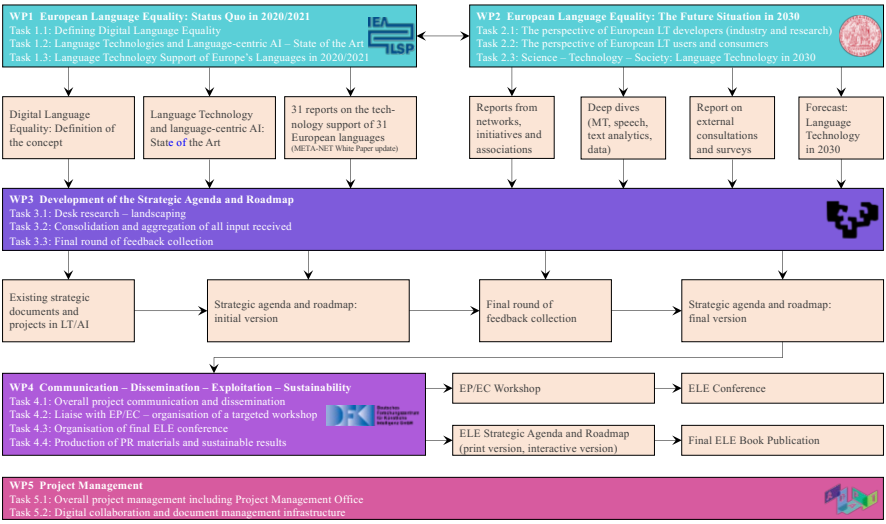

Fig. 2 Work packages and tasks of the ELE project

severe limitations (Bender et al. 2021). Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models and self-supervised systems opens up the way to leverage LT for less-developed languages. For the first time, a single multilingual model recently outperformed the best specially trained bilingual models on news translations, i. e., one multilingual model provided the best translations for both low- and high-resource languages, indicating that the multilingual approach appears to be the future of MT (Tran et al. 2021). However, the development of these new systems would not be possible without sufficient resources (experts, data, compute facilities, etc.), including the creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application.

Unfortunately, as of now, there is no equality in terms of tool, resource and application availability across languages and domains. Although LT has the potential to overcome the linguistic divide in the digital sphere, most languages are neglected for various reasons, including an absence of institutional engagement from decision-makers and policy stakeholders, limited commercial interest and insufficient resources. For instance, Joshi et al. (2020) and Blasi et al. (2022) look at the relation between the types of languages, resources and their representation in NLP conferences over time. As expected, but also disappointingly, only a very small number of the over 6,000 languages of the world are represented in the rapidly evolving field of LT. A growing concern is that due to unequal access to digital resources and financial support, only a small group of large enterprises and elite universities are in a position to lead further development in this area (Ahmed and Wahed 2020).

To unleash the full potential of LT in Europe and ensure that no users of these technologies are disadvantaged in the digital sphere *simply due to the language they*

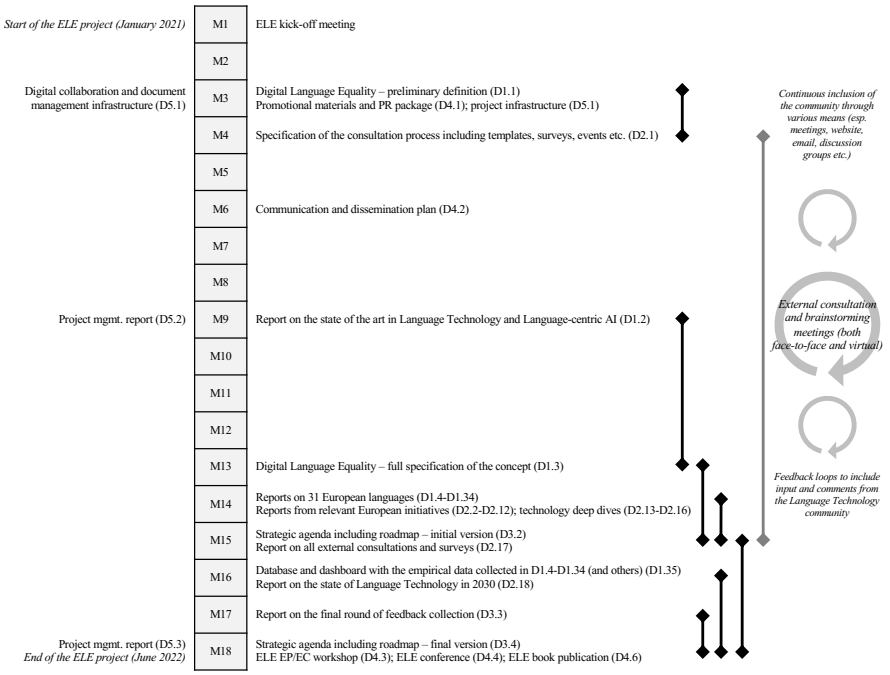

Fig. 3 Overall timeline of the ELE project

speak, we argue that there is a pressing need to facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical, fair and unbiased language understanding and communication. In short, we must ensure DLE in all areas of society, from government to business to citizens.

4 Summary of this Book

This book is structured into two main parts. Part I examines the *current state of play* of technology support for Europe's languages. Part II outlines the *future situation* in 2030 and beyond, as specified through the community consulting and forecasting process of the ELE project. Below we include short summaries of the two parts.

4.1 Part I: European Language Equality – Status Quo in 2022

Part I concentrates on the current situation as of 2022. First, Chapter 2 examines the state-of-the-art in LT, NLP and language-centric AI. It provides the technical foundation of all subsequent chapters. Chapter 3 defines the DLE metric, developed

within the project, with its technological (Gaspari et al. 2022) and contextual factors (Grützner-Zahn and Rehm 2022). This chapter also describes the interactive DLE dashboard, which was implemented as an additional component of the European Language Grid cloud platform (ELG, Rehm 2023). Assuming that the ELG catalogue of resources, tools and services contains, at any given point in time, a representative picture of the technology support of Europe's languages, the dashboard can be used to visualise the overall situation in different ways, including comparisons of multiple languages along various dimensions. Chapter 4 summarises the findings and provides an answer to the question of how Europe's languages compare technologically ca. 2022. The chapter describes the methodology of basing the computation of the DLE scores on the contents of the ELG repository, which has been substantially expanded by the ELE project with more than 6,000 additional resources, and highlights the current situation using a number of graphs. Chapters 5 to 37 contain extended high-level summaries of the 33 language reports produced by the ELE project. These reports can be conceptualised as updates, ten years on, of the META-NET White Papers (Rehm and Uszkoreit 2012), especially as many of them were written by the original authors.

4.2 Part II: European Language Equality – The Future Situation in 2030 and beyond

Part II outlines the future situation in 2030 and beyond, making use of the collected and synthesised results of the community consultation process. First, Chapter 38 describes the community consultation process on a general level, primarily with regard to the different surveys used in the project vis-à-vis European LT developers, European LT users and consumers as well as European citizens. The chapter also summarises the approach regarding the four technology deep dives as well as the dissemination and feedback collection activities in the project. Chapter 39 summarises the results of the three main surveys. The following four chapters highlight the main findings of the four technology deep dives on the four main areas of LT research and development: Machine Translation (Chapter 40), Speech Technologies (Chapter 41), Text Analytics (Chapter 42) as well as Data and Knowledge (Chapter 43). The penultimate Chapter 44 presents the strategic plans and projects in LT and AI from an international, European and national perspective. It contextualises the strategic recommendations of the project. Finally, Chapter 45, provides an extended summary of the stand-alone document of the *Strategic Research, Innovation and Implementation Agenda and Roadmap* the ELE project has developed.[10] On the whole, the present book can be conceptualised as the collective findings and recommendations of the ELE project, and as such it reflects years of work based on the distilled input and collaboration of hundreds of experts and stakeholders from across the European LT and language-centric AI community.

[10] https://european-language-equality.eu/agenda/

References

Ahmed, Nur and Muntasir Wahed (2020). "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research". In: *CoRR* abs/2010.15581. https://arxiv.org/abs/2010.15581.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada, pp. 610–623.

Blasi, Damian, Antonios Anastasopoulos, and Graham Neubig (2022). "Systematic Inequalities in Language Technology Performance across the World's Languages". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5486–5505. DOI: 10.18653/v1/2022.acl-long.376. https://aclanthology.org/2022.acl-long.376.

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)*. 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. 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. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.2.pdf.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. https://aclanthology.org/2020.acl-main.560.

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

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Federico Gaspari, German Rigau, Maria Giagkou, Stelios Piperidis, Annika Grützner-Zahn, Natalia Resende, Jan Hajic, and Andy Way (2022). "The European Language Equality Project: Enabling digital language equality for all European languages by 2030". In: *The Role of National Language Institutions in the Digital Age – Contributions to the EFNIL Conference 2021 in Cavtat*. Ed. by Željko Jozić and Sabine Kirchmeier. Budapest, Hungary: Nyelvtudományi Kutatóközpont, Hungarian Research Centre for Linguistics, pp. 17–47.

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. https://aclanthology.org/L18-1519.pdf.

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 etc.: Springer. http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

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. https://data.europa.eu/doi/10.2861/136527.

Tran, Chau, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan (2021). "Facebook AI's WMT21 News Translation Task Submission". In: *Proceedings of the Sixth Conference on Machine Translation (WMT 2021)*. Online: Association for Computational Linguistics, pp. 205–215. https://aclanthology.org/2021.wmt-1.19.