EUROPEAN LANGUAGE EQUALITY 2

D4.2

Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap – Supplementary Document

Authors	ELE 2 Consortium
Dissemination level	Public
Date	30-06-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
Deliverable number	D4.2
Deliverable title	Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap – Supplementary Document
Type	Report
Number of pages	16
Status and version	Final
Dissemination level	Public
Date of delivery	30-06-2023
Work package	WP4: Dissemination, Communication and Stakeholder Engagement
Task	Task 4.4 Publication of revised strategic agenda (online)
Authors	ELE 2 Consortium
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Acronyms

ADA	Audio Data Augmentation
BLARK	Basic LAnguage Resource Kit
CLARIN	Common Language Resources and Technology Infrastructure
DLE	Digital Language Equality
EDIC	European Digital Infrastructure Consortium
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality
ELE 1	European Language Equality (preceding project)
ELE 2	European Language Equality *(this project)*
ELEN	European Language Equality Network
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resources Association
ELSNET	European Network of Excellence in Language and Speech
EU	European Union
FSTP	Financial Support to Third Parties
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HPC	High-Performance Computing
HPC	High Performance Computing
LR	Language Resource
LT	Language Technology/Technologies
LLM	Large Language Models
META-NET	EU Network of Excellence to foster META
NLP	Natural Language Processing
SLU	Spoken Language Understanding
SME	Small and Medium-sized Enterprises
SRIA	Strategic Research and Innovation Agenda
STOA	Science and Technology Options Assessment
SWOT	Strengths, Weaknesses, Opportunities and Threats

Abstract

The Strategic Research and Innovation Agenda (SRIA) was developed by the European Language Equality (ELE) initiative to address the imbalance of technology support among European languages. The SRIA provides a roadmap and priority research themes to guide Europe in its future advancements in language technology and promote inclusivity in the digital sphere. As a project *from* the community *for* the community, various stakeholders, including companies, academic organisations, language institutes, funding agencies and policymakers were actively involved in shaping the SRIA. Feedback was collected through various rounds of consultations, projects and dissemination activities during the runtime of ELE 1 (01/2021-06/2022) and ELE 2 (07/2022-06/2023). The most recent version of the SRIA is openly accessible as the book *European Language Equality – A Strategic Agenda for Digital Language Equality* (Rehm and Way, 2023a), especially Chapter 45 (Rehm and Way, 2023b). This supplementary document, Deliverable D4.2, provides additional insights, recommendations and revisions compiled and collected by the project ELE 2.

1 Introduction

Despite the findings regarding digital language inequality presented in the META-NET White Paper Series more than ten years ago (Rehm and Uszkoreit, 2012), there still remains a significant imbalance in the level of technology support for the languages of Europe. While English benefits from a wide range of resources and technologies, the majority of other languages face a substantial lack of technological support. The results obtained in the ELE 1 project further emphasise that there is still a stark imbalance in terms of technology support among European languages in 2022/2023.

In ELE 1, we developed the first version of the Strategic Research and Innovation Agenda for Digital Language Equality in Europe by 2030 (SRIA). The SRIA addresses the imbalance of technology support in a comprehensive way and offers concrete recommendations to establish full digital language equality in 2030. One of the goals of the ELE 2 project has been to revise, extend and further promote the SRIA and the included roadmap. For this purpose, additional feedback from relevant stakeholder groups has been reviewed and consolidated throughout the project. The research partners liaised with national and international funding agencies as well as policymakers and partners from industry and research to ensure that the SRIA meets all expectations set by these groups.

To further enhance the visibility of the SRIA, the consortium called for support from the community to endorse the SRIA and provide more feedback, especially on the recommendations that present a substantial part of the SRIA. In addition, the consortium launched an open call for SRIA contribution projects, financially supported through the Financial Support to Third Parties mechanism (FSTP). We supported a total of nine projects that produced clearly defined use cases and best-practice examples of language resource development and language technology implementations in relevant industry sectors and other areas.

The ELE initiative successfully finalised the initial version of the SRIA and presented Version 1.0 in November 2022 at the STOA Workshop (see below). The book *European Language Equality – A Strategic Agenda for Digital Language Equality*, published in June 2023, includes an extended summary of the SRIA. This version of the SRIA had already undergone multiple rounds of reviews and incorporated substantial feedback from the community and all stakeholders. Thus, Deliverable D4.2 presents additional feedback we collected through consultation rounds, the contribution projects and the SRIA endorsement initiative in 2023. The collected feedback strongly supports the current version of the SRIA.

2 Status and Structure of SRIA

The overall vision and unifying goal of the ELE Programme is to achieve complete digital language equality in Europe by 2030. The programme was prepared jointly with many relevant stakeholders from the European Language Technology (LT), Natural Language Processing (NLP), Computational Linguistics and language-centric AI communities, as well as with representatives of relevant initiatives and associations, and language communities. While the ELE 1 project (01/2021-06/2022) laid most of the groundwork for the development of the SRIA, the ELE 2 consortium added many important contributions. Table 1 lists the different versions of the SRIA, including their publishing dates.

Version 0.9	Available on the ELE website (early November 2022)
Version 1.0	Available on the ELE website (late November 2022)
Version 1.5	The ELE book (Rehm and Way, 2023a), especially Chapter 45 (finalised in March 2023, published in June 2023)
Version 2.0	Version 1.5 plus this supplementary document (Deliverable D4.2) available on the ELE website (June 2023)

Table 1: History of the different SRIA versions

Version 0.9 of the SRIA was published on November 7 of 2022 and presented in a STOA workshop in the European Parliament on 8 November 2022.[1]

Figure 1: STOA Workshop “Language Equality in the Digital Age” (November 2022)

This event was the third in a series of STOA workshops on language technologies (the first was held in 2013 and the second in 2017). It built on the STOA study *Language Equality in*

[1] Further details are available in Rehm et al. (2022). Information and materials related to this STOA Workshop, including recordings of the sessions, are available online at https://www.europarl.europa.eu/stoa/en/events/details/towards-full-digital-language-equality-i/20220711WKS04301. In addition, third-party sources also covered this high-profile workshop, for instance the popular online journal Slator dedicated an article to the event: https://slator.com/european-parliament-should-support-digital-language-equality.

the Digital Age, which led to the European Parliament's resolution (European Parliament, 2018) of the same name. The STOA event aimed to explore the research and development environment of language technologies in the context of EU multilingualism and presented the project results. A panel discussion allowed policymakers and experts from academia and industry to discuss challenges and opportunities for digital language equality in the EU.

The updated version 1.0 of the SRIA was published on 19 November 2022 (Consortium, 2022). The SRIA 1.0 Version was divided into the following eight chapters:

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. 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
8. Concluding Remarks

Since late November 2022, the SRIA had been reviewed and extended in various regards. Version 1.5 was finalised in March 2022 and is available in the form of the recently published book *European Language Equality – A Strategic Agenda for Digital Language Equality* (Rehm and Way, 2023b). While the whole book outlines the current state of Language Technology in Europe and gives recommendations, Chapter 45 (Rehm and Way, 2023b) includes the revised and extended summary of the ELE Strategic Research and Innovation Agenda.

The 30-page book chapter explores the relationship between Europe's multilingualism and the importance of striving for digital language equality. It discusses the challenges and opportunities presented by linguistic diversity in the digital sphere and tackles how language equality can solve it. Further, the recommendations, that serve as pillars for the shared ELE Programme are discussed. To conclude the chapter, the main components of the roadmap along with the needed actions, budget, timeline and collaborations are presented. Figure 3 shows the table of contents of Chapter 45 called "Strategic Research, Innovation and Implementation Agenda for Digital Language Equality in Europe by 2030".

3 SRIA Stakeholder Consultations

The SRIA has always been intended to be a community project, which included an active dialogue with all stakeholders involved. ELE 2 continued the work carried out in ELE 1 to strengthen the engagement of stakeholders who can contribute to or benefit from the developed agenda. The various versions presented in Section 2 made use of the additional feedback collected throughout the two projects.

Work Package 1 of the ELE 2 project focused on documenting additional consultations with various stakeholders, such as policymakers, funding agencies, and language institutes. Its main objectives were to recognise obstacles that hinder institutes from expanding and exchanging resources, advocate for the ELE initiative, and assess the level of dedication towards establishing the comprehensive ELE Programme. This task involved facilitating the endorsement of the SRIA and documenting stakeholder backing for the recommendations put forth in the SRIA (Hegele et al., 2022; Kirchmeier et al., 2022; Gaspari et al., 2023).

Figure 2: Book *European Language Equality* and Deliverable D3.4

Figure 3: ELE Book, Chapter 45 (extended summary of SRIA) – Table of Contents

Policymakers

Although we have received very positive feedback from policymakers in the European Parliament (for example, in the STOA Workshop, see above) and the European Commission regarding the findings and strategic recommendations of the ELE initiative, there has not been any definite commitment made regarding the financing and implementation of the ELE Programme at the time of writing (June 2023).

With regard to the European Commission, representatives of the consortium met with Mariya Gabriel, Commissioner for Innovation, Research, Culture, Education and Youth in March 2023. The meeting, facilitated by Jordi Solé (MEP), was positive and a concrete follow-up step was suggested by Commissioner Gabriel. Briefly after this meeting, Mariya Gabriel left the EC to take up a leadership role in Bulgaria on the way to form a new government.

During the runtime of ELE 2, various developments happened in parallel, including the funding and establishment of the Common European Language Data Space[2] and the emergence of the Language European Digital Infrastructure Consortium (EDIC). Possible interactions and concrete synergies between these newly established initiatives are currently under discussion and will be determined in the autumn of 2023 (Gaspari et al., 2023).

Language Institutes

The two main associations representing European language communities that have been consulted during the SRIA development process are EFNIL and ELEN, both are members of the ELE 2 consortium. EFNIL[3] (European Federation of National Institutions for Language) aims to promote linguistic diversity and multilingualism in Europe, bringing together national institutions and organisations responsible for language planning, policy, and research in European countries. Similarly, ELEN[4] (European Language Equality Network) advocates for linguistic diversity and language rights in Europe by focusing on supporting minority and regional languages, promoting language learning opportunities, and raising awareness about language diversity issues within the European Union. EFNIL and ELEN consulted with their members. Below, we present some of the collected findings (Kirchmeier et al., 2022).

Most institutes that maintain language resources are generally open to sharing them for the betterment of their respective languages. Many have already shared their resources through infrastructures like ELG and CLARIN. However, there are legal barriers, such as copyright and GDPR, that hinder further sharing of resources. This is especially true for older resources that were compiled at a time when the significance of language resources for the advancement of language technology and artificial intelligence was not as widely recognised as it is today.

The SRIA discusses the availability of data and knowledge resources and the limitations posed by GDPR restrictions. The European Language Technology SWOT Analysis, that the consortium conducted, clearly labels GDPR and copyright as a weakness that needs to be addressed. It is important that GDPR and also IPR regulations become more flexible. Clear legal national frameworks and efficient transpositions and implementation of the European directives on open data are essential to ensure well-regulated access to language data for research and innovation (non-commercial and commercial) purposes.

The dialogue with the language institutes has been ongoing and has led to many fruitful discussions. The conference META-FORUM 2023 titled “Digital Language Equality for a Multilingual Europe”[5], taking place in Brussels on June 27, will discuss these topics further in a dedicated session. This session, “The Role of National and Regional Language Institutes

[2] https://digital-strategy.ec.europa.eu/en/funding/language-data-space-call-tenders
[3] http://www.efnil.org
[4] https://elen.ngo
[5] https://european-language-equality.eu/meta-forum-2023/programme/

for Digital Language Equality" will be chaired by Sabine Kirchmeier, president of the European Federation of National Institutions for Language, and bring together the following panellists: Frieda Steurs (Instituut voor de Nederlandse Taal, Netherlands), Kristine Eide (Norwegian Language Council, Norway), Svetla Koeva (Bulgarian Academy of Sciences, Bulgaria), Rickard Domeij (Language Council of Sweden, Sweden) and Davyth Hicks (European Language Equality Network).

4 SRIA Contribution Projects: New and Revised Recommendations

In order to incorporate external ideas and expertise into the SRIA, the ELE 2 consortium initiated an open invitation for SRIA Contribution Projects, making use of the Financial Support to Third Parties (FSTP) mechanism.

The ELE 2 FSTP Project Board, in collaboration with the wider ELE 2 consortium, identified several areas within the SRIA that could benefit from an expansion. They put forth a list of ten topics for which applicants could submit proposals through the open call. Detailed descriptions of each topic can be found in Deliverable D3.4 (Aldabe et al., 2023). Based on evaluations, the FSTP Project Board selected nine projects, estimated to run for a duration of three months, from January to March 2023. Ultimately, the selected projects, consisting of three from companies and six from research organisations, contributed to the topics listed below.

Topic 1: Data Sets for More Robust Speech Technology

One of the SRIA recommendations for all LT research areas is to gather and make available the necessary critical mass of resources in terms of data. The goal is to create sufficient multilingual and multimodal data of quality (responsible, legal, diverse, unbiased, ethical, representative, etc.), in all European languages and domains (media, health, legal, etc.).

Speech is one of the priority research themes that are being suggested as part of the ELE Programme. The current Speech recommendations of the SRIA include enhancing speech resources and acoustic models for diverse languages, developing natural synthetic voices, improving context modeling, handling challenging audio conditions, integrating speech with other modalities, and addressing privacy and security concerns.

The selected SRIA contribution projects strongly support the SRIA research recommendations related to Speech.

- The project **NGT-Dutch Hotel Review Corpus** conducted by Tilburg University in the Netherlands provided high-quality data valuable for the development of sign language technology. Additionally, as an outcome of the project, a multimodal parallel corpus has been created, containing both Dutch and the Sign Language of the Netherlands (NGT). The availability of such a corpus can support research focused on creating more inclusive language technology, incorporating sign language technology into the broader language technology landscape and making it an integral part of LT development.
- The project **Building E2E Spoken-Language Understanding Systems for Virtual Assistants in Low-Resources Scenarios** by Balidea in Spain includes the collection of over 250 hours of recordings involving more than 11,000 participants. The dataset is likely to be one of the largest in Spoken Language Understanding (SLU) to date. Additionally, the project developed a voice collection tool and real-time data dashboards. The project also provided recommendations to guide the design and collection of datasets

for end-to-end (E2E) SLU systems. Furthermore, it established guidelines for approaching E2E SLU projects in low-resource scenarios, leveraging Balidea's expertise in e-Health virtual assistants.

- The project **Multilingual and Mixed Language Data for Inclusive Speech Technology** carried out by Ghent University in Belgium offers recommendations to address language inequalities in current speech and language technology. It highlights that existing technologies are predominantly built with monolingual assumptions and fail to account for social, linguistic, and cultural factors, among speakers and users. Consequently, there is a shortage of multilingual and mixed language data crucial for developing inclusive speech technologies in Europe. By emphasising the need to bridge this gap, the project aims to foster advancements that promote linguistic equality and inclusivity in speech technology.
- The project **Generation of a Large Speech Corpus for Spain Languages using Data Augmentation** conducted by Pangeanic in Spain has developed a methodology and set of guidelines for conducting an audio project with augmentation. They collected text segments, audio files and corresponding metadata for each language, resulting in datasets consisting of audio files with metadata. Additionally, they applied audio data augmentation techniques to a subset of the data and provided a sample of 8,459 audio files in Asturian with audio data augmentation applied.
- The project **Underrepresented Speech Dataset from Open Data: Case Study on the Romanian Language** conducted by the Research Institute for Artificial Intelligence in Romania identified multimedia platforms that offer content under open licenses, specifically Creative Commons licenses. They downloaded relevant samples of open multimedia content and annotated them with metadata, including information such as the number of speakers, age, gender and speech quality. Additionally, the multimedia files were transcribed, resulting in aligned text with the speech. The project includes annotation guidelines that are useful for constructing similar datasets in other languages. These guidelines can serve as a valuable resource for researchers and practitioners working on building speech datasets in different linguistic contexts.

Topic 5: General NLP/LT Domains (Desk Research)

General NLP/LT systems often struggle when presented with data from domains or topics they have not been specifically trained on. The lack of coverage in particular domains poses restrictions for advancements in LT in these areas, e. g., medical, health, pharmaceutical, legal, finance, insurance, science, manufacturing, publishing etc. Adapting models and techniques to effectively handle out-of-domain data and generalise well across various domains remains a challenge. Research is thus needed to find faster, cheaper, more reliable and if possible multilingual methods and procedures that will generate the necessary datasets in a short time and in good quality.

One of the SRIA recommendations for all LT research areas is:

To develop better benchmarks and datasets (ethical, responsible, legal, etc.) for all languages, domains, tasks and modalities.

The development of new LT systems would not be possible without sufficient resources (data, experts, computing facilities, etc.). The creation of carefully designed and constructed evaluation benchmarks and annotated data sets for every language and domain of application is needed to foster technological progress, while encouraging a deeper understanding of the mechanisms by which they are achieved.

- The project **European LT Domains** conducted by the University of Zagreb analysed data on domains, NLP/LT tasks, and languages, focusing on a set of 39 different European languages. An overview of LT usage across different domains has been provided, along with a detailed examination of languages, domains, and NLP tasks. The study identified fragmentation within the European LT community, showing that certain domains receive more attention. To address this, it was recommended to gain a better understanding of highly sought-after domains and leverage insights from actively explored NLP areas to support underdeveloped domains.

Topic 7: Computing facilities for LT (Desk Research)

The SRIA infrastructure recommendations strongly suggest distributed research infrastructures and flexible access to sufficient HPC facilities.

To ensure flexible access to GPU-based HPC facilities and a more suitable computing infrastructure.

Distributed research infrastructures and flexible access to sufficient HPC facilities are important because they promote collaboration, optimise resource utilisation, provide scalability, enhance accessibility and inclusivity, and accelerate scientific discoveries. They empower researchers to leverage high-performance computing capabilities and advance their research endeavours.

The following contribution project emphasises the importance of this recommendation.

- The project **Computing Facilities for LT** conducted by the University of Zagreb included the preparation of a comprehensive list of 56 manually curated High-Performance Computing (HPC) systems in Europe, providing data on various aspects. Additionally, a survey was conducted that shed light on real-world HPC usage and needs. Based on the findings, a recommendation was made to ensure that the current AI HPC infrastructure, particularly advanced GPUs, offers sufficient, flexible, and dynamic access policies and protocols. This is necessary to continuously support advanced deep learning technologies for all types of users, including public and private researchers and developers.

Topic 10. Basic LAnguage Resource Kit (BLARK) (Desk Research)

The BLARK[6] concept was devised more than 20 years ago in a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) and first launched with a Dutch initiative called Dutch Human Language Technologies Platform that was initiated in April 1999. The BLARK defines, ideally in a language-independent way, the minimal set of language resources to do any precompetitive language and speech technology research at all for a language (Krauwer, 1998).

The following SRIA policy recommendation strongly supports an adaptation of BLARK.

To define and develop a BLARK-like minimum set of language resources and capacities that all European languages should possess.

[6] http://www.blark.org

- The project **A BLARK for Minority Languages in the Era of Deep Learning: Expertise from Academia and Industry** by imaxin|software in Spain developed an analytical method based on existing BLARKs for languages like Dutch and Faroese. They created a BLARK matrix for minority languages, providing researchers with a comprehensive tool to evaluate their development. The project recommends using the expertise of the Nós Project and imaxin|software to make the Galician case a global example for language development, particularly for languages mentioned in the SRIA.
- The project **Artificial Intelligence Data Kit 2030** conducted by the Institute for the Bulgarian Language included the latest advancements and trends in large language models (LLMs). Additionally, surveys were conducted to identify and assess the most significant datasets and benchmarks used for training and evaluating LLMs. The project also involved an overall analysis and specification of the AI Data Kit. Furthermore, a database of over 200 AI companies from across Europe was compiled. The companies were asked to complete a brief questionnaire regarding their involvement in the development, adaptation, and use of AI applications. Based on the findings, the project recommended that a single static universal kit of text, audio, image, and video data is not feasible in the modern context of rapid technological development and varying levels of technological support for different European languages.

5 SRIA Endorsement: New and Revised Recommendations

To streamline the process of endorsing the SRIA and to effectively capture the support and commitment of stakeholders towards its recommendations, we have developed a web form and made it available on the ELE website[7]. The web form serves as a platform to introduce the SRIA, including a link to access the complete document. Stakeholders were requested to indicate their endorsement of the SRIA. Additionally, they had the option to focus on specific recommendations and offer their supplementary feedback for each group of recommendations. Deliverable D3.4 contains the full feedback collected through the endorsement form, as well as additional information on the participants' demographics and organisation types (Aldabe et al., 2023).

The SRIA breaks down the concrete recommendations for the shared programme. First, possible cornerstones for policy recommendations are outlined, as well as ideas for the realisation of a governance model. Second, the technology and data recommendations suggested by the ELE consortium are revised as well as the suitable infrastructure. Further, research recommendations are considered ground-breaking and game-changing by the LT community. Over the last decade, the community has developed a clear vision of the work needed in the different areas of LT. The European Parliament has also acted on these ideas.

These SRIA recommendations aim to advance digital language equality, deepen natural language understanding, and address language barriers across Europe.

Policy Recommendations

The policy recommendations' main objectives include establishing a large-scale funding programme for LT research and education, protecting regional and minority languages, recognising the rights of linguistic minorities in the digital world, promoting mother tongue teaching, ensuring sufficient funding and resources for LT development, fostering collaboration between academia and industry, defining essential language resources and capacities, facilitating language data sharing, enabling access to computing infrastructure and sensitive

[7] https://european-language-equality.eu/endorse-the-ele-sria/

data processing, supporting SMEs and startups in utilizing LT, attracting and retaining international LT professionals and achieving European LT sovereignty.

The following slight extensions were suggested:

Language Communities

Language communities should decide the extent of LT use for their languages, as technologies like MT are sometimes perceived as threatening to indigenous and minority languages.

The voice of the language communities could be added to this SRIA recommendation:

> To ensure comprehensive EU-level legal protection for the more than 60 regional and minority languages.

Scope of Language Equality

Within-language disparities in digital inclusion can arise due to the complexity of digital texts, resulting in certain groups of speakers being excluded. To address this issue, the proposed ELE programme aims to foster inclusion by eliminating comprehension and readability barriers for non-expert readers, thereby promoting equal access to digital information and services.

This disparity could be emphasised in the following SRIA recommendation:

> To empower recognition of the collective rights of national and linguistic minorities in the digital world (including sign languages).

Indigenous Languages

A proposal has been made to explicitly include the term “indigenous languages” alongside “lesser-used, regional, and minority languages” in the SRIA.

Indigenous languages could be added to this SRIA recommendation:

> To ensure comprehensive EU-level legal protection for the more than 60 regional and minority languages.

Further, the current set of policy recommendations could be extended by at least one recommendation that focuses clearly and explicitly on PR activities to raise awareness.

The feedback we received included:

- Further support for awareness-raising activities targeting national governments so that they invest in LT for their own languages.
- Given the social and economic impact of digital language equality, the policy of “LT as a free service” has been advocated.

Technology and Data Recommendations

The Technology and Data Recommendations in the SRIA focus on advancing technology, data utilisation, and language resources. Key points include developing high-performance applications, addressing data gaps, promoting open ecosystems, overcoming data inequality,

leveraging public sector data, and supporting green language technologies. Other recommendations include fostering collaboration between research and industry, improving access to multilingual online services, and preventing digital extinction by defining minimum language resources.

Research in Rule-Based, Knowledge-Based and Hybrid Approaches

A future ELE Programme should avoid exclusive focus on the machine learning paradigm and reserve, instead, some funding for research in rule-based, knowledge-based and hybrid approaches as these may prove more effective for less-spoken languages where the small amounts of data that are available may not be sufficient to adequately support machine learning techniques.

If not listed as a separate recommendation, it could further support the point of unequal data availability.

To develop methods to overcome the unequal data availability, by focusing on, e. g., annotation transfer, multilingual models preserving quality, few-shot or zero-shot learning.

Language Communities

Language communities have deep knowledge and expertise in their respective languages. Involving them more would ensure that the data and resources created for European languages are culturally sensitive, respectful, and representative of the diverse linguistic and cultural landscapes in Europe. Hence, their active involvement could be further highlighted in the SRIA and expand the following recommendation.

To address the lack of available data and define the minimum of language resources and capacities that all European languages should possess.

6 Conclusions

The ELE Strategic Research and Innovation Agenda (SRIA) aims to address the existing imbalance in technology support for different languages in Europe and establish full digital language equality by 2030. The SRIA was developed through extensive collaborations with stakeholders, including large enterprises and SME companies, academic organisations, language institutes, policymakers and funding agencies. The SRIA underwent multiple rounds of reviews and revisions to gather feedback and incorporate input from the community. The engagement of stakeholders was crucial to ensure the relevance and effectiveness of the agenda. Policymakers, such as the European Parliament and the European Commission, expressed positive feedback but have not made definitive commitments regarding financing and implementation. The language communities, represented through associations such as EFNIL and ELEN, played an active role in the SRIA consultations, providing valuable insights and resources.

To further enrich the SRIA, the ELE 2 consortium initiated an open call for SRIA Contribution Projects. The nine selected projects focused on critical areas identified in the agenda. The projects contributed valuable datasets, guidelines, and further recommendations that help advance language technology and promote linguistic equality and inclusivity.

The final SRIA and its consultation process demonstrate the commitment of the ELE initiative and stakeholders to bridge the technological gap among European languages. The community has agreed on a comprehensive roadmap with strategic recommendations to guide

research, innovation, and implementation efforts in the field of language technology. Challenges related to data availability, legal frameworks, and resource sharing remain. However, the ELE initiative aims to create a shared European Programme that benefits all languages and ensures linguistic diversity.

References

Itziar Aldabe, Aritz Farwell, Maria Giagkou, Jan Hajic, Jana Hamrlova, Stelios Piperidis, and German Rigau. Deliverable D3.4 Consolidation and curation of all input and feedback received, 2023. URL https://european-language-equality.eu/wp-content/uploads/2023/06/ELE2___Deliverable_D3_4.pdf. Project deliverable; EU project European Language Equality 2 (ELE 2); Grant Agreement no. 01884166 – 101075356 ELE.

ELE Consortium. Deliverable D3.4 Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap, 2022. URL https://european-language-equality.eu/reports/SRIA-and-roadmap.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

European Parliament. 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, 2018.

Federico Gaspari, Jane Dunne, Maja Popović, Andy Way, Maria Giagkou, Stelios Piperidis, Jana Hamrlová, Davyth Hicks, Sabine Kirchmeier, Katrin Marheinecke, and Georg Rehm. Deliverable D1.3 Report on all consultations with stakeholders, 2023. URL https://european-language-equality.eu/deliverables_ELE2/D1_3_consultations-all-stakeholders.pdf. Project deliverable; EU project European Language Equality 2 (ELE 2); Grant Agreement no. 01884166 – 101075356 ELE.

Stefanie Hegele, Annika Grützner-Zahn, Katrin Marheinecke, Georg Rehm, Maria Giagkou, and Stelios Piperidis. Deliverable D1.1 Specification of approach for consultations and for documentation of stakeholder commitment, 2022. URL https://european-language-equality.eu/deliverables_ELE2/D1_1_consultation-and-stakeholder-commitment.pdf. Project deliverable; EU project European Language Equality 2 (ELE 2); Grant Agreement no. 01884166 – 101075356 ELE.

Sabine Kirchmeier, Georg Rehm, Marie Mattson, Davyth Hicks, and Jane Dunne. Deliverable D1.2 Report on consultations with funding agencies, policy makers and language institutes, 2022. URL https://european-language-equality.eu/deliverables_ELE2/D1_2_consultations-funding-agencies.pdf. Project deliverable; EU project European Language Equality 2 (ELE 2); Grant Agreement no. 01884166 – 101075356 ELE.

Steven Krauwer. Elsnet and elra: A common past and a common future. *ELRA Newsletter*, 3(2):4–5, 1998.

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

Georg Rehm and Andy Way, editors. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer, 6 2023a. doi: https://doi.org/10.1007/978-3-031-28819-7.

Georg Rehm and Andy Way. Strategic Research, Innovation and Implementation Agenda for Digital Language Equality in Europe by 2030. In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 387–412. Springer, Cham, Switzerland, 6 2023b. doi: https://doi.org/10.1007/978-3-031-28819-7_45.

Georg Rehm, Stefanie Hegele, and Katrin Marheinecke. Deliverable D4.3 Report on EP/EC workshop, 2022. URL https://european-language-equality.eu/reports/EC-workshop.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.