



# EUROPEAN <sup>2</sup> LANGUAGE EQUALITY

## D3.4

### Consolidation and curation of all input and feedback received

---

Authors	Itziar Aldabe (UPV/EHU), Aritz Farwell (UPV/EHU), Maria Giagkou (ILSP), Jan Hajic (CUNI), Jana Hamrova (CUNI), Stelios Piperidis (ILSP), German Rigau (UPV/EHU)
Dissemination level	Public
Date	02-06-2023

---

## 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	D3.4
Deliverable title	Consolidation and curation of all input and feedback received
Type	Report
Number of pages	39
Status and version	Final
Dissemination level	Public
Date of delivery	02-06-2023
Work package	WP3: Strategic Research, Innovation & Deployment Agenda: Maintenance and Extension
Task	Task 3.4 Collection and curation of all input and feedback received
Authors	Itziar Aldabe (UPV/EHU), Aritz Farwell (UPV/EHU), Maria Giagkou (ILSP), Jan Hajic (CUNI), Jana Hamrlova (CUNI), Stelios Piperidis (ILSP), German Rigau (UPV/EHU)
Reviewers	Sabine Kirchmeier (EFNIL), Federico Gaspari (DCU)
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 <a href="http://www.european-language-equality.eu">http://www.european-language-equality.eu</a>  © 2023 ELE2 Consortium

## 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

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Further Consultations: Stakeholder Commitment</b>	<b>1</b>
2.1	Consultations: Policymakers, Funding Agencies, Language Institutes etc. . . . .	1
2.2	SRIA endorsement and stakeholder commitment . . . . .	2
<b>3</b>	<b>Contribution Projects</b>	<b>6</b>
3.1	FSTP . . . . .	7
3.2	Topic 1. Data sets for more robust speech technology . . . . .	7
3.2.1	Project NGT-Dutch Hotel Review Corpus (Tilburg University, Netherlands)	7
3.2.2	Project Building E2E spoken-language understanding systems for virtual assistants in low-resources scenarios (Balidea, Spain) . . . . .	8
3.2.3	Project Multilingual and Mixed Language Data for Inclusive Speech Technology (Ghent University, Belgium) . . . . .	9
3.2.4	Project Generation of a large speech corpus for Spain languages using Data Augmentation (Pangeanic, Spain) . . . . .	9
3.2.5	Project Underrepresented speech dataset from open data: case study on the Romanian language (Research Institute for Artificial Intelligence, Romanian Academy, Romania) . . . . .	10
3.3	Topic 5. General NLP/LT Domains (Desk Research) . . . . .	10
3.3.1	Project European LT Domains 2023 (University of Zagreb, Faculty of Humanities and Social Sciences, Croatia) . . . . .	10
3.4	Topic 7. Computing facilities for LT (Desk Research) . . . . .	11
3.4.1	Project Computing facilities for LT (University of Zagreb, Faculty of Humanities and Social Sciences, Croatia) . . . . .	11
3.5	Topic 10. Basic LAnguage Resource Kit (BLARK) (Desk Research) . . . . .	12
3.5.1	Project A BLARK for minority languages in the era of deep learning: expertise from academia and industry (Factoría de Software e Multimedia, S.L. (imaxin software), Spain) . . . . .	12
3.5.2	Project Artificial Intelligence Data Kit 2030 (Institut for Bulgarian Language Prof. Lyubomir Andreychin, Bulgaria) . . . . .	12
<b>4</b>	<b>Maintenance and Extension</b>	<b>13</b>
4.1	Strategic Documents . . . . .	13
4.2	Missing Resources and Stakeholders . . . . .	14
4.3	Results . . . . .	15
<b>5</b>	<b>Conclusions</b>	<b>16</b>
<b>A</b>	<b>Appendix: Full list of unique organisations that filled in the endorsement form</b>	<b>19</b>
<b>B</b>	<b>Appendix: All feedback collected through the online endorsement form</b>	<b>28</b>

## List of Figures

1	Types of organisations that endorsed the SRIA . . . . .	3
2	The SRIA endorsement form as published online (page 1/4) . . . . .	20
3	The SRIA endorsement form as published online (page 2/4) . . . . .	21
4	The SRIA endorsement form as published online (page 3/4) . . . . .	22
5	The SRIA endorsement form as published online (page 4/4) . . . . .	23

## List of Tables

1	Number of stakeholders who filled in the online endorsement form per country	4
---	--	---

## List of Acronyms

AI	Artificial Intelligence
CF	Contextual Factors
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
ELE1	European Language Equality (preceding project)
ELE2	European Language Equality ( <i>this project</i> )
ELEN	European Language Equality Network
ELG	European Language Grid (EU project, 2019-2022)
EU	European Union
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
MT	Machine Translation
NLP	Natural Language Processing
TF	Technological Factors

## Abstract

ELE 2 continues to review key research areas and gaps in research that need to be addressed to ensure that the current inequality in LT support for Europe's languages can be overcome. This deliverable reports on material prepared in WP1 (Further Consultations and Documentation of Stakeholder Commitment), WP2 (Open Call for Contribution Projects), and WP3 (Maintenance and Extension). It provides a synthesis of 1) the input and feedback received from consultations with stakeholders, policymakers, and language institutes, 2) the results obtained from the SRIA contribution projects, 3) the outcomes derived from an analysis of LT and language-centric AI reports, policies, and initiatives, and 4) the conclusions drawn from a prioritized list of missing language resources and stakeholders.

## 1 Introduction

This deliverable consolidates and curates material prepared in WP1, WP2, and WP3 so that it may be included in Deliverable 4.2, the “Strategic agenda and roadmap”. WP1 was dedicated to reporting on further consultations with stakeholders, including policymakers, funding agencies, and language institutes. Its underlying goals were to identify roadblocks that prevent institutes from developing and sharing greater resources, to promote the ELE initiative, and to gauge commitment towards the establishment of a large-scale ELE programme. Part of the work involved facilitating endorsement of the SRIA and documenting stakeholder support for the SRIA recommendations. WP2 was devoted to the organization and implementation of an open call for SRIA contribution projects through the Financial Support for Third Parties (FSTP) mechanism. Its purpose was to incorporate additional external ideas and expertise from several fundamental areas into the SRIA. The selected projects contributed to four topics: datasets for more robust speech technology, general NLP/LT domains, computing facilities for LT, and a Basic Language Resource Kit (BLARK). WP3 was responsible for the maintenance and extension of the ELE SRIA. Its objectives were to monitor LT and language-centric AI reports, policies, and initiatives, manage and update the ELE dashboard, and specify a prioritized list of missing language resources and stakeholders.<sup>1</sup> An overview of the results generated by these work packages is provided in the pages that follow.

## 2 Further Consultations: Stakeholder Commitment

### 2.1 Consultations: Policymakers, Funding Agencies, Language Institutes etc.

It is envisaged that achieving digital language equality (DLE) in Europe by 2030 can only be accomplished through the support and commitment of multiple stakeholder groups working together towards the goal of establishing a joint large-scale programme. ELE 2 partners engaged with the relevant policymaking bodies and funding agencies, research, industry as well as consumer and user stakeholders at European, national and regional levels to systematically expand the list of stakeholders identified in ELE 1 and document their commitment. The activities undertaken to reach relevant policymakers, funding agencies, language institutes and language communities, including the methodology applied (e. g. organisation of questionnaires, direct consultations, follow-up interviews, etc.), are described in Kirchmeier et al. (2023).

---

<sup>1</sup> The update to the ELE dashboard is presented in detail in D3.3.

We have received positive feedback from policymakers, such as the European Parliament and the European Commission, with regard to the ELE initiative's findings and strategic recommendations, but at the time of writing no concrete commitment concerning the financing and implementation of the ELE Programme. However, other developments have taken place in parallel, including the funding and establishment of the Common European Language Data Space<sup>2</sup> and the emerging Language European Digital Infrastructure Consortium (EDIC). It remains to be determined how these various newly created initiatives will eventually work together and what the concrete synergies between them will be in the future (see section 2.2 of Gaspari et al. (2023a)).

The consultation with institutes for national languages (EFNIL) and minority languages (ELEN) was designed, on the one hand, to promote the ELE initiative and, on the other, to identify missing resources and issues that prevent the institutes from developing and sharing greater resources. It demonstrated that the institutes primarily compile mono- and bilingual dictionaries, corpora and terminology resources, but also construct the tools that are needed to search and process these resources. Furthermore, institutes for minority languages in particular also produce technological resources such as machine translation (MT) and speech synthesis or voice recognition systems because these are generally not offered by tech companies.

For the most part, all institutes that maintain resources are willing to share whatever possible for the benefit of their languages. Many do so via ELG, CLARIN or other platforms. The main impediment to sharing even greater resources are legal issues such as copyright and GDPR. This is especially the case for older resources that were compiled in times where the awareness of the usefulness of language resources for the development of LT and AI was not as high as it is today. Some of these restricted resources are available for research purposes, but not for public or industrial use. Another obstacle is the lack of experts with knowledge of both the language and the technology.

A considerable amount of institutes would like to have access to more text data, audio data and historical texts. In addition, they require access to conversational AI and spoken dialogue systems to enable speakers of their languages to participate in the digital age. It is important to keep in mind that languages with a significant number of speakers and a governmental support structure, whether national or regional, such as Catalan (10 million speakers), possess greater resources and more advanced technology than those with few speakers, resources, and access to only basic technology, such as Cornish (1000 speakers). There is consensus that the most beneficial initiatives to improve this situation would involve funding for acutely under-resourced languages, a workaround for the legal issues, and more training for researchers and developers with deep knowledge of these languages.

## 2.2 SRIA endorsement and stakeholder commitment

To facilitate the endorsement of the SRIA and to additionally document stakeholders' support of and commitment to the SRIA recommendations, we created a web form and published it on the ELE website.<sup>3</sup> The form presented an introduction to the SRIA with an appropriate link to the full document. The main groups of recommendations (i.e. policy, governance, technology, data, research and implementation recommendations) were summarised to a concise paragraph. The stakeholders (organisations and experts from the field) were thus requested to state their support for the whole SRIA and, optionally, to concentrate on specific recommendations and provide their additional feedback per recommendations group.

All stakeholders reached through WP1 Tasks T1.2 and T1.3 were requested to visit the endorsement form and to provide their feedback for each group of recommendations. The

<sup>2</sup> <https://digital-strategy.ec.europa.eu/en/funding/language-data-space-call-tenders>

<sup>3</sup> <https://european-language-equality.eu/endorse-the-ele-sria/>

form was additionally publicised through all appropriate ELE 2 communication channels, i.e. the ELT Newsletter, Twitter account, etc.<sup>4</sup> In total 254 individuals endorsed the SRIA through the online form (until 17/05/2023). They come from 40 different countries, mostly European, but notably also some non-European ones (e.g. South Africa, Peru and USA). The distribution of respondents per country is presented in Table 1.

In dedicated questions, the respondents were asked to name their organisation(s) and indicate its type. Note that the former was a free-text question and the latter was multiple-choice, as a single respondent may be affiliated with more than one organisation.

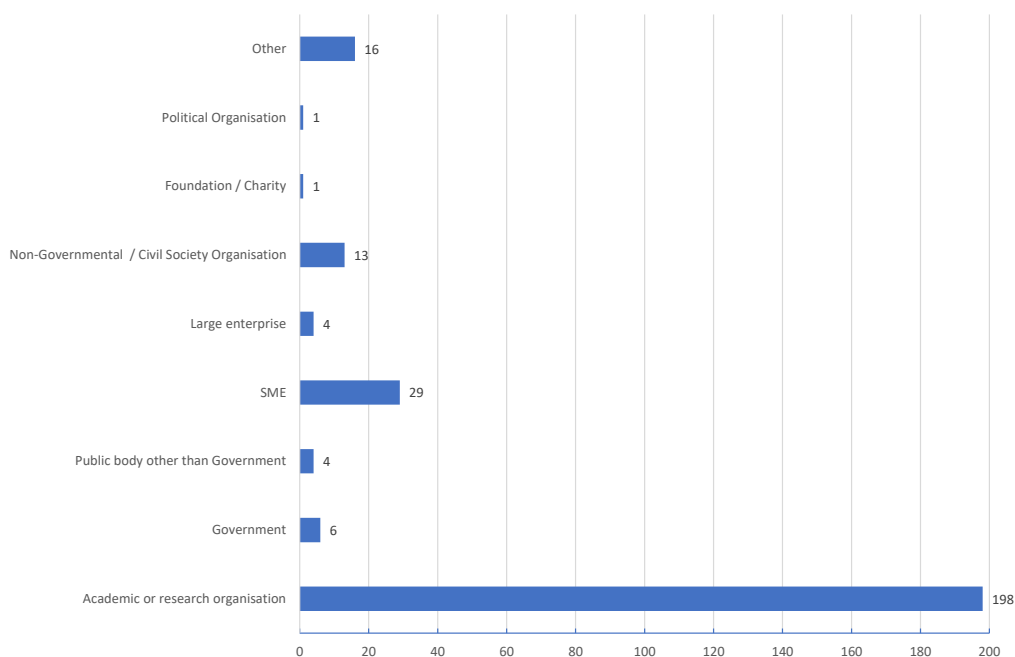


Figure 1: Types of organisations that endorsed the SRIA

According to the answers to these questions, the respondents represent 163 different organisations (see full list of organisations in Appendix A). Most of them are from research and academia (Figure 1). The second most populated type of organisations are SMEs (29), suggesting that, even though smaller languages may not constitute a lucrative market segment, the European SMEs that are active in the field do recognise the importance of DLE. Some representatives of European governments also filled in the endorsement form, despite the fact that the ELE 2 consortium decided not to focus (at the time) on this type of stakeholders because of the ongoing efforts to establish a Language EDIC. The governmental bodies that filled in the endorsement form are:

- the Ministry of Culture of Slovakia
- the Secretary for Language Policy, Government of Catalonia
- the National Council of the Maltese Language
- the State Language Department of Georgia
- the Center for the Luxembourgish Language

<sup>4</sup> <https://european-language-equality.eu/endorse-the-ele-sria/>

<b>Country</b>	<b>Number of stakeholders</b>
Spain	69
Czechia	19
Germany	15
UK	13
France	11
Ireland	11
Italy	10
Netherlands	9
Romania	8
Greece	8
Belgium	7
Austria	7
Denmark	7
Poland	6
Malta	4
Latvia	4
Finland	3
Sweden	3
Croatia	3
Norway	3
Luxembourg	3
Serbia	3
Georgia	2
Iceland	2
Bulgaria	2
Switzerland	2
Slovakia	2
Slovenia	2
Hungary	2
Estonia	2
Lithuania	2
Cyprus	1
South Africa	1
Republic of Mali	1
United Arab Emirates	1
Moldova	1
Peru	1
Faroe Islands	1
Portugal	1
USA	1
n/a	1
<b>Total</b>	<b>254</b>

Table 1: Number of stakeholders who filled in the online endorsement form per country



The NGOs/civil society organizations/charity foundations that largely represent the language communities of regional and minoritised languages have been relatively active in stating their support through the online form. Some of them include:

- Association of Translation Companies
- European Language Equality Network
- Tallers per la Llengua
- ISSA Polska - Association for Information Systems Security
- Plataforma per la Llengua (Catalonia)
- Croatian Language Technologies Society
- Kamusi Project International and African Academy of Languages (ACALAN)
- Cambra d'Òc
- Afûk
- Impact Nation Institute
- European Federation of National Institutions for Language (EFNIL)
- Network to Promote Linguistic Diversity

The full feedback collected through the endorsement form is presented in Appendix B. In summary, leaving aside various points that are already covered in the SRIA, the following additional arguments and suggestions have been collected through the endorsement form with respect to **policy** recommendations:

- The role of **education** has been underlined. In particular, mother-tongue teaching should be supported.
- Some respondents have argued that language communities should decide the extent of LT use for their languages, as technologies like MT are sometimes perceived as threatening for indigenous and minority languages.
- The scope of language equality has been extended by a respondent from a cross-languages issue to a within-language one. Groups of speakers of the same language may be excluded from the digital sphere because of the level of complexity of digital texts. It is thus suggested that an ELE programme should promote inclusion by removing **comprehension and readability** barriers for non-expert readers.
- 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.
- For reasons of completeness, it has been suggested that the term “indigenous languages” should be explicitly added to the SRIA together with “lesser-used, regional and minority languages”.

- **International cooperation** should be sought in order to promote DLE for all of the world's languages.<sup>5</sup>

The following points were proposed with respect to **data and technology** recommendations:

- 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 available data will never adequately support machine learning techniques.
- Ensure that all languages have access to basic LT (e. g. keyboards, proofing tools, etc.) and allow language communities to take over localisation of OS systems independently of software makers.
- Language communities should be considered critical stakeholders when it comes to collecting data for smaller languages, e. g. through crowd-sourcing.
- As already proposed in the SRIA, respondents argued that the development of **open-source large language models** for all EU languages is an essential starting point for reengineering all traditional NLP tasks.

The notion of openness has been iteratively emphasised in the collected feedback. One of the respondents focused in particular on the Open Data Directive, which must be expanded to include language data as high value datasets. In addition, it was recommended that the ELE programme should intensively promote the idea of fair use of data to exonerate researchers from the threat of IPR violations.

The governance recommendations were well received, notably the proposal for allocating the area of multilingualism, linguistic diversity and LT to the portfolio of an EU Commissioner. Some respondents additionally suggested a more active involvement in the governance scheme of **industrial research labs** and stakeholders from **Digital Humanities**, e. g. university libraries.

Finally, with respect to the implementation of a future ELE programme, it has been argued that apart from EC and national funding, especially through multi-country projects, funding opportunities from the European Defense Fund should be investigated. Similarly, involvement and investment from industry must also be guaranteed.

### 3 Contribution Projects

To bring external ideas and expertise to the SRIA in a substantial way, the ELE 2 consortium organized an open call for SRIA Contribution Projects utilising the Financial Support to Third Parties (FSTP) mechanism. These projects, of an expected duration of approximately three months, were meant to provide meaningful and compelling input for the strategic agenda and roadmap in the form of project reports. They were also intended to potentially generate software, datasets, studies and analyses that might be used for the strategic agenda as well as for other communication and dissemination purposes, depending on the nature of the proposed work.<sup>6</sup>

<sup>5</sup> A specific recommendation and proposal for cooperation has come from the 55 Member States of the African Union, since *“the needs and rationales for African language equality are exactly the same as those for European language equality, and the methods deployed to reach linguistic equity in Europe are largely similar to those proposed by Africa’s designated intergovernmental language organ”*.

<sup>6</sup> D2.1, “Open Call Setup and Results”, contains details on the FSTP projects.

### 3.1 FSTP

The FSTP Project Board, together with the broader ELE 2 consortium, mapped crucial areas where the SRIA needs to be extended or updated and proposed 10 topics to which applicants to the open call could submit their proposals (a detailed description of individual topics is available in the Call Documentation on the ELE 2 website):

1. Data sets for more robust speech technology
2. Study of language coverage for text mining and natural language understanding in key European industrial sectors
3. Legal Assessment (Desk Research)
4. General NLP/LT/AI Landscaping (Desk Research)
5. General NLP/LT Domains (Desk Research)
6. Analysis of AI and LT in European news media
7. Computing facilities for LT (Desk Research)
8. Demonstrably Greener Models of MT
9. Survey of the use of LT in the hospital sector
10. Basic Language Resource Kit (BLARK) (Desk Research)

The call was opened from September 29 to November 29, 2022 and 36 proposals from 24 different applicants were eligible for full evaluation. Applicants could submit more proposals, but a maximum of one proposal for each topic was set. The proposals were submitted to 9 different topics (no proposal was submitted for topic 3 - Legal Assessment). The FSTP Project Board selected 9 projects with a total grant amount of 185,000 EUR. The estimated duration of the projects was 3 months, from January to March 2023. During the selection process, the FTSP Project Board attempted, in accordance with the evaluation criteria, to select at least one project for each topic. However, due to the small number of proposals submitted to several of the topics, it was not possible to cover them all. In the end, 9 selected projects (3 from companies, 6 from research organisations) contributed to 4 of the topics. The next sections present the projects' main results and the recommendations <sup>7</sup>

### 3.2 Topic 1. Data sets for more robust speech technology

#### 3.2.1 Project NGT-Dutch Hotel Review Corpus (Tilburg University, Netherlands)

The goal of this project was to create a multimodal parallel corpus of Dutch and Sign Language of the Netherlands (NGT).<sup>8</sup> The corpus contains hotel reviews written in Dutch and their translation into NGT (videos) provided by 6 deaf professional NGT sign language translators. It consists of 283 reviews and a total of 21,825 words. The NGT translations constitute almost 4 hours of videos (213.18 minutes). The duration of the NGT videos ranged from around 10 seconds to around 4 minutes.

<sup>7</sup> Also see <https://european-language-equality.eu/open-call>

<sup>8</sup> The parallel corpus is available at the ELG platform <https://live.european-language-grid.eu/catalogue/corpus/21566>. Find more information about the corpus and its creation in the project report: [https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2\\_Project\\_Report\\_NGT\\_HoReCo.pdf](https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2_Project_Report_NGT_HoReCo.pdf)

## Results

- quality data for the development of sign LT
- multimodal parallel corpus of Dutch and Sign Language of the Netherlands (NGT)

## Recommendations

The availability of such a corpus supports research focusing on more inclusive LT and contributes in particular to the efforts towards making sign LT part of the equation.

### 3.2.2 Project Building E2E spoken-language understanding systems for virtual assistants in low-resources scenarios (Balidea, Spain)

This project carried out a study on the minimum design features of a spoken language understanding (SLU) dataset for low-resource scenarios with a high presence of linguistic variety (considering that Galician is less-resourced than Spanish). The project proposed quality measures, regardless of the language of application, to determine the complexity of the designed dataset in order to establish minimums in the design and collection of data. An ambitious campaign was designed and executed to collect voices in Galician.<sup>9</sup> The effort represents an unprecedented milestone for the language, involving the development and updating of a voice collection tool, dashboards with real-time data, and the creation of digital content on the internet, among other activities.<sup>10</sup>

## Results

- more than 250 hours of recordings of more than 11,000 participants covering 98% of Galicia
- perhaps the largest SLU dataset obtained to date<sup>11</sup>
- development and update of a voice collection tool, dashboards with real-time data, and the creation of digital content on the internet<sup>12</sup>

## Recommendations

This project aims to contribute to the SRIA by presenting guidance for designing and collecting datasets for end-to-end (E2E) spoken language understanding (SLU) systems, establishing guidelines on how to approach an E2E SLU project in a low-resource scenario, taking advantage from Balidea's experience in e-Health virtual assistants.

<sup>9</sup> Real-time data from the campaign (which is still active), can be viewed at: <https://falai.balidea.com/datos-temporal/public/dashboard/c2d965e3-287d-4187-a95e-482fbe82578f/>

<sup>10</sup> Detailed results of the data collection campaigns, methods of validation for the collected data, and quality control measurements of the final dataset may be found in the project report: [https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2\\_Project\\_Report\\_BEST\\_Assistants.pdf](https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2_Project_Report_BEST_Assistants.pdf)

<sup>11</sup> The designed dataset is in the database in the json format. A xlsx format with the dataset information can be found at: [https://docs.google.com/spreadsheets/d/1zmnwCB7XzWkKsv7Ej\\_fYIthrTn79YRWif/edit?usp=sharing&ouid](https://docs.google.com/spreadsheets/d/1zmnwCB7XzWkKsv7Ej_fYIthrTn79YRWif/edit?usp=sharing&ouid)

<sup>12</sup> The recording tool is accessible via: <https://falai.balidea.com/>. Examples of the content created for the campaigns may be viewed at: <https://youtu.be/gE8E-yHMveE>; <https://youtu.be/wJtKf-xR43A>; [https://www.instagram.com/reel/Cp93hj5riSp?utm\\_source=ig\\_web\\_copy\\_link](https://www.instagram.com/reel/Cp93hj5riSp?utm_source=ig_web_copy_link); and [https://www.instagram.com/reel/CpFTwdeoF56/?utm\\_source=ig\\_web\\_copy\\_link](https://www.instagram.com/reel/CpFTwdeoF56/?utm_source=ig_web_copy_link)

### 3.2.3 Project Multilingual and Mixed Language Data for Inclusive Speech Technology (Ghent University, Belgium)

Focusing on the younger members of an underrepresented immigrant community in Belgium, the project collected multilingual speech data (Turkish-Dutch-English) based on natural conversations and transcribed them by developing its own guidelines, which will be shared with the community to serve as an example for future studies.

#### Results

- At the time of writing, the project's results have not yet been delivered. It will complete its execution on May 31<sup>st</sup>, 2023.

#### Recommendations

Current speech and LT are built with monolingual assumptions ignoring the variation (e.g., social, linguistic and cultural) among different types of speakers/users that creates language inequalities. There is a lack of multilingual and mixed language data to build these technologies in Europe.

### 3.2.4 Project Generation of a large speech corpus for Spain languages using Data Augmentation (Pangeanic, Spain)

This project involved collecting audio data in 4 different languages spoken in Spain by recruiting participants to record their voices. To increase the size of the dataset, Audio Data Augmentation techniques were applied.<sup>13</sup>

#### Results

- developed a methodology for conducting an audio project with augmentation and created a set of guidelines outlining the necessary steps
- collected text segments and audio files for each language, along with their corresponding metadata
- created datasets consisting of 63,271 audio files with metadata totaling approximately 180 hours of content<sup>14</sup>
- provided the corresponding code and applied audio data augmentation to a subset of the data
- provided a sample of 8,459 audio files in Asturian with ADA applied

<sup>13</sup> The process is described in detail in the project report: [https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2\\_Project\\_Report\\_SpeechCorpus.pdf](https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2_Project_Report_SpeechCorpus.pdf).

<sup>14</sup> All the datasets are available on the ELG platform: speech and transcription corpus in Catalan <https://live.european-language-grid.eu/catalogue/corpus/21545>; speech and transcription corpus in Basque <https://live.european-language-grid.eu/catalogue/corpus/21546>; speech and transcription corpus in Galician <https://live.european-language-grid.eu/catalogue/corpus/21547>; speech and transcription corpus in Asturian <https://live.european-language-grid.eu/catalogue/corpus/21534>; and speech and transcription Corpus with Augmented Audio Data in the Asturian Language <https://live.european-language-grid.eu/catalogue/corpus/21548>

## Recommendations

To contribute to the ELE 2 strategic agenda, this project creates a guideline for building an extensive speech dataset with transcriptions of languages spoken in Spain through audio data augmentation (ADA) techniques. By incorporating ADA techniques, such as noisy backgrounds, time masking, and speed variation, it aims to expand a standard dataset into a much larger speech dataset by a factor of at least 20. This approach provides a better representation of various speech and sound types, as it simulates real-life environments.

### 3.2.5 Project Underrepresented speech dataset from open data: case study on the Romanian language (Research Institute for Artificial Intelligence, Romanian Academy, Romania)

This project studied the usability of open data for building speech datasets for types of voices that are usually missing or underrepresented in existing speech datasets.<sup>15</sup> It identified existing multimedia open data, including platforms, types of media, percent of usable voices in a data sample, types of open licenses, types of underrepresented voices (including children, young people, older people, women, etc.), and percent of underrepresented voices. In a second step, it conducted a case study on the Romanian language and the same methodology may be applied to any other language.

## Results

- identified multimedia platforms offering content under open licenses (especially Creative Commons licenses)
- downloaded relevant samples of open multimedia content and annotated them with metadata, including number of speakers, age, gender and speech quality<sup>16</sup>
- multimedia files were transcribed, producing text aligned with the speech

## Recommendations

The annotation guidelines (useful for building similar datasets in other languages) are available as an Annex in the project report.

## 3.3 Topic 5. General NLP/LT Domains (Desk Research)

### 3.3.1 Project European LT Domains 2023 (University of Zagreb, Faculty of Humanities and Social Sciences, Croatia)

The objective of this project was to provide a snapshot illustrating how NLP/LT are utilised in various sectors and to report on the fields that make frequent use of NLP/LT. The study identified domains that employ active NLP/LT techniques and domains that require special and additional attention.

<sup>15</sup> The project website is available at: <https://www.racai.ro/p/uspdatro/>. The project report, including the revised version of the annotation guidelines, will be available on the ELE 2 website <https://european-language-equality.eu/open-call/>

<sup>16</sup> The annotated files were assembled into a dataset, including the metadata. The audio files were segmented, producing speech files with the corresponding text transcription. The dataset was released on Zenodo, European Language Grid and the RELATE platform: <https://doi.org/10.5281/zenodo.7898232>; <https://live.european-language-grid.eu/catalogue/corpus/21567>; and <https://relate.racai.ro/repository/uspdatro>.

## Results

- compilation and analysis of collected data about domains, NLP/LT tasks and languages<sup>17</sup>
- overview of LT usage by different domains regarding a language set composed of 39 European languages as defined in the first ELE project.
- overview regarding each dimension of this study (languages, domains, and NLP tasks)
- heat maps that detail the usage of the selected NLP tasks by the different domains were generated for each language. This language-specific examination was completed with a general overview of the distribution of LT by domain.

## Recommendations

The fundamental fragmentation that exists among the LT community in Europe favours some domains over the others. The essential action to help the underdeveloped or non-developed NLP/LT domains can be taken with a better insight into the existing highly desired domains. An overview of all the actively explored NLP areas leads to a better understanding of underdeveloped domain-specific characteristics.

## 3.4 Topic 7. Computing facilities for LT (Desk Research)

### 3.4.1 Project Computing facilities for LT (University of Zagreb, Faculty of Humanities and Social Sciences, Croatia)

This project provides a feasibility analysis of existing HPC services in Europe in terms of their current support for LT. The objective of the project was to evaluate HPC and, particularly, a GPU's capacity for both small and large trials, the available access protocols and their compatibility, the GPU's capacity for training large neural models, etc. The project focuses on the analysis of various HPC configurations and the enumeration of details regarding various user-important aspects. These include the performance of the HPC, the number of GPUs, access for various user types, access modes, etc.<sup>18</sup>

## Results

- detailed list of 56 manually curated HPC systems in Europe containing data for the previously described aspects
- 26 responses were obtained from the survey conducted to investigate the aspects of real-world HPC usage and needs
- video summarising the findings of desk research and a survey.

---

<sup>17</sup> Mutual relationships between members of these three sets were presented in the project report: [https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2\\_Project\\_Report\\_EuLTDom2023.pdf](https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2_Project_Report_EuLTDom2023.pdf). The extracted data per language (CSV files) and the graphs in the SVG format are also available in the digital material accompanying the project report: <https://github.com/dfvalio/EuLTDom2023>.

<sup>18</sup> Further details may be found in the project report: [https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2\\_Project\\_Report\\_ComFac4LT.pdf](https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2_Project_Report_ComFac4LT.pdf).

## Recommendations

Current AI HPC infrastructure in the form of advanced Graphic Processing Units (GPUs) should provide sufficient, flexible and dynamic access policies and protocols to provide continuous support to advanced deep learning technologies to all kind of users including public and private researchers and developers.

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

#### 3.5.1 Project A BLARK for minority languages in the era of deep learning: expertise from academia and industry (Factoría de Software e Multimedia, S.L. (imaxin | software), Spain)

This project created a BLARK that recognizes and analyzes the key resources and tools necessary to develop state-of-the-art LT. In order to develop this project, researchers focused on the Galician case from two perspectives: industry (imaxin | software) and academy (Nós Project). The methodology employed had three main steps: first, previous BLARK proposals were analyzed in depth and the previously proposed criteria that are still prevailing in the deep learning era, versus those that are outdated, were identified. Second, available resources were surveyed for different European languages in order to separate the minimum requirements from the desirable maximums. Third, the project team deployed its own BLARK using its knowledge of Galician to test its capacity to appropriately produce a realistic and informative panorama of its degree of development.<sup>19</sup>

## Results

- taking existing BLARKs as a starting point, such as those for Dutch or Faroese, the project team developed an analytical method to determine the basic requirements and tools for any language to get closer to the state of the art in LT
- desk research in which a BLARK matrix to minority languages was elaborated. With this tool, LT researchers and experts now possess a wide-covering and flexible BLARK to evaluate the development of minority languages.

## Recommendations

Combining the research experience of the Nós Project in the collection of data for Galician and the creation of models with the experience of imaxin | software in the development of tools for minority languages for end users, the Galician case will be an example for any language in the world and, in particular, for those presented in the SRIA.

#### 3.5.2 Project Artificial Intelligence Data Kit 2030 (Institut for Bulgarian Language Prof. Lyubomir Andreychin, Bulgaria)

The main objective of this project was to specify the data kit (corpora, models, datasets, etc.) required to develop computer applications classified as artificial intelligence. Drawing on an in-depth analysis of existing studies, the project proposed an AI data kit for language understanding, generation, and transformation, as well as a set of criteria to which the data kit

---

<sup>19</sup> The full project report is available at: [https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2\\_Project\\_Report\\_BLARK.pdf](https://european-language-equality.eu/wp-content/uploads/2023/04/ELE2_Project_Report_BLARK.pdf).



was adapted depending on technological advancement and the specific technology support for different languages.<sup>20</sup>

## Results

- a survey of the most recent and notable large language models (LLMs), demonstrating trends and advancements
- surveys on the most significant datasets and benchmarks currently available for LLMs training and evaluation
- overall analysis and specification of the AI Data Kit
- compiled a database of over 200 AI companies from across Europe and asked them to complete a brief questionnaire on their involvement in the development, adaptation, and use of AI applications.<sup>21</sup>

## Recommendations

In the modern conditions of: a) rapid technological development; b) varying degrees of technological support for different European languages, it is not viable to suggest a single static universal kit of text, audio, image, and video data.

# 4 Maintenance and Extension

## 4.1 Strategic Documents

Our review of strategic documents and projects analysed international, national and regional Strategic Research Agendas, studies, and initiatives related to LT and language-centric artificial intelligence (Aldabe et al., 2022). This synopsis of international and European reports included a SWOT analysis designed to identify factors that must be addressed to help solve the pressing issue of digital language inequality in Europe. As in the case of the original ELE SRIA, the resulting recommendations from this endeavour, catalogued below, will be incorporated into the revised ELE SRIA. Given AI's and LT's growing technological significance, it should come as no surprise that the documents surveyed reflect the attention paid to AI and LT across social, political, and economic domains. In addition to trends in innovation, many discuss the socioeconomic and political impact of AI and LT from a policy perspective, including the question of multilingualism and equal technological support for Europe's languages through the application of LT.

Several of the reports we reviewed agree that LT is one of three core and rapidly growing application areas within AI, together with vision and robotics. This is partly because LT has become the nerve center of the software that processes unstructured information and exploits the vast amount of data contained in text, audio and video files, including those from the web and social media. In fact, various consulting firms forecast significant growth in the global LT market based on the explosion of applications observed in recent years and the expected exponential growth in unstructured digital data. The field's brisk development and increasing social relevance is highlighted in national and regional AI and LT strategies both inside and outside of Europe, as well as in prioritized areas for research, development

<sup>20</sup> The report "FSTP Project Report AID 2030 – Artificial Intelligence Data Kit 2030" will be published soon on the ELE 2 website. A 3-minute presentation of the project is available at: <https://we.tl/t-Z06sAIvEcQ>.

<sup>21</sup> The survey findings are included in the project report as Appendix 2 and the survey is available at: <https://surveyplanet.com/p2l65zu1>.

and innovation. Emphasis is placed, in particular, on its importance as an economic driver and as a possible tool to provide solutions for societal challenges. The perspective goes hand in hand with the observation that as AI has left the laboratory it has quickly demonstrated a significant “real-world impact on people, institutions, and culture”.<sup>22</sup> While EU citizens generally view this ongoing development in a positive light, concern over the socioeconomic, legal and ethical impact of AI is warranted. Trustworthy AI that respects European values is considered essential in the efforts to strengthen European AI and digital sovereignty.

It is within this greater context that the issue of European multilingualism generally surfaces in reports. It is a topic that is often viewed through the prism of the problems associated with language barriers and the awareness that many European languages are endangered or on the edge of extinction. It is an unfortunate truth that, although official EU languages are granted equal status politically, they are far from equally supported technologically.<sup>23</sup> And while it may be that no silver bullet exists to remedy the situation, language-centric AI affords a means to provide immediate support. It can not only help bypass many of the obstacles currently standing in the way of better cross-language and cross-border European communication, economic growth and social stability, but also alleviate some of the pressures on languages that are in danger of digital extinction (Moseley, 2010; Rehm and Uszkoreit, 2012; STOA, 2017; European Parliament, 2018). The importance of doing so cannot be overstated. As has been argued elsewhere, sophisticated multilingual, cross-lingual and monolingual LT for all European languages would future-proof our languages as cornerstones of our cultural heritage and richness. In this vision, yet to be fully realized, LT is made in Europe for Europe, tailored to meet the unique cultural, social, and economic demands of its citizenry. Powerful homegrown LT would offer a mechanism to turn a linguistically fragmented Europe into a truly unified and inclusive one. To reach this point, the EU’s decisive role in building Europe’s digital society must be met half way by the member states, who will need to do their part to enact the individual AI and LT strategies that are best suited to address their particular demands.

## 4.2 Missing Resources and Stakeholders

In deliverable 3.2, “Missing Resources and Stakeholders” (Aldabe et al., 2023), we utilised the newly updated version of the ELE dashboard and drew from the grassroots knowledge contained in each of the 35 language reports prepared for the ELE project (Rehm and Way, 2023) to determine which language resources are the most underrepresented and whether potential stakeholder types remain to be consulted. ELE’s interactive dashboard, one of the project’s central contributions, provides a mechanism for dynamically exposing and monitoring the support languages receive through LT. As detailed in Gaspari et al. (2022, 2023b), it is based on the ELG database and provides an overview of the DLE metric, which includes technological factors (TFs) and contextual factors (CFs). The recent updates to the dashboard, in the form of heatmaps, radial bars, and evolution over time provide visualisations that allow for novel cross-language comparisons. The first displays the actual number of resources or percentages representing the contribution of each language per resource type. The second enables users to create charts that display data using a circular layout. The third permits users to generate charts that display overall data evolution over time or the intensity at which it progressed over time. The data we were able to analyse through the dashboard helped us identify the main language resource types that are currently lacking across all languages and expose gaps that must be filled in the future.

Along with the development of the DLE metric, one of the principal aims of the ELE project has been to provide a sounding board for the diverse members of the LT community. Our

<sup>22</sup> <https://ai100.stanford.edu>

<sup>23</sup> See the DLE Dashboard <https://live.european-language-grid.eu/catalogue/dashboard>

belief is that digital language equality can only be attained by taking as many voices from this community as possible into account. As part of this effort, we set out to ensure that no significant stakeholders had been left out of the initial consultation process. To do so, we reviewed ELE's language reports to find stakeholder types that may not have been approached. The effort involved all partners in the ELE consortium and yielded an overall picture of the types of stakeholders that are involved in LT and language-centric AI across Europe. Through this collaborative exercise, we were able to compile a list of stakeholder types that helped us establish which have been approached and which might still require greater attention.

### 4.3 Results

The results of D3.1 are reflected in a set of six recommendations, gathered under the rubrics of policy, governance model, technology and data, infrastructure, research, and implementation (Aldabe et al., 2022). Together, these recommendations are designed to guide the construction of a shared large-scale programme that will ensure European digital language equality by 2030. This collaborative programme must refocus and strengthen European LT and NLP research through a pan-European effort. The top-line findings are:

- (1) Such a programme must involve all stakeholders. This includes the full participation of research centres, academia, SMEs, startups and others. But, just as importantly, it will also require a coordinated and resolute effort between the EU and participating countries and regions that includes adequate policymaking, distributed research infrastructures and technological platforms. Failure to establish this multilayered coordination will leave in place a status quo in which language communities are too small or fragmented to fully protect the digital future of their languages.
- (2) Interdisciplinary research will be fundamental to the success of the programme as LT is aggregated and applied to more complex settings. Greater basic and applied research is needed for language data collection (text, dialog, vision, sign language and other forms of interactions), speech analysis, AI, human-computer interaction, machine learning, robotics, and tasks such as machine reading, text analysis, MT, chatbots, virtual assistants and summarisation.
- (3) Funding and further investment are needed at all levels. EU funding will enable overarching coordination and an EU-wide technological infrastructure. National and regional funding must complement EU funding with regard to language-specific research and development. If insufficient investment in the underdeveloped areas of LT and language-centric AI persists, the result might be the digital extinction of various neglected European languages.

In a nutshell, any successful large-scale initiative to address the problem of digital language inequality in Europe must be built upon the three core pillars of cooperation, research, and funding at every level.

The results of D3.2 are divided into those for currently inadequate language resources and those for potentially missing stakeholders. With respect to the former, the top three language resource categories that consistently demonstrate the weakest support across the majority of languages in terms of absolute numbers are 1) image and video processing, 2) human-computer interaction, and 3) grammar. Moreover, while it is true that each language differs in terms of language resource distribution, a substantial number of languages possess no, or close to no, resources for some categories of language resources. This alarming state of affairs is compounded by the inescapable observation that the level of language support rapidly deteriorates as one moves from the official EU languages to national and regional co-official languages to those with no official recognition at all. Thus, although there is some

consistency across all language groups, our main finding is that each language requires specialised attention to address its particular needs. Experts in a determined language must decide which language resource categories should be prioritised and efforts to give languages a sufficient technological foundation must go hand in hand with the visions and strategies of stakeholders across all sectors.

With regard to the review of possible missing stakeholders, relevant stakeholders mentioned in the language reports were broadly categorised into the following spheres: government, industry, research institutions, and independent organisations. The diverse stakeholders grouped within these overlapping spheres belong to fairly well-defined subsets and, given the similarity between types of LT stakeholders that exist across most languages, they are generally consistent. Taking this into account, we found that the ELE project has consulted nearly every stakeholder type at some level to build a full picture at all relevant levels. By way of example, the research institution, industry, and government spheres all have a significant presence in the language reports. The former two have been consulted in some depth through various surveys conducted by ELE (Thönnissen, 2022; Eskevich and de Jong, 2022; Rufener and Wacker, 2022; Hajič et al., 2022; Hegele et al., 2022; Gísladóttir, 2022; Blake, 2022; Hrasnica, 2022; Heuschkel, 2022; Way et al., 2022).

## 5 Conclusions

As we have seen, the consultations, feedback, and information gathered by ELE have returned a host of results and recommendations that will be used to refine and finalise the SRIA and Roadmap. First and foremost is the strong belief that attaining digital language equality in Europe by 2030 will require a commitment from multiple stakeholders working under the auspices of a collaborative large-scale programme. The results of D3.1, reflected in a set of six recommendations, are designed to guide the construction of this programme, which must ensure that European LT and NLP research is strengthened through a pan-European effort. Any successful large-scale initiative to address the problem of digital language inequality in Europe must be built upon cooperation, research, and funding at every level. Policymakers, including the European Parliament and the European Commission, have offered positive feedback concerning ELE's findings and strategic recommendations, but have not committed as yet to the implementation and financing of a future ELE Programme.

In the meantime, several obstacles, proposals, and types of initiatives are consistently singled out as essential to either overcome or mitigate the problem of digital language inequality. A handful concern legal issues and greater training for researchers and developers, especially for under-resourced languages. Institutes that are dedicated to the study and protection of national and minority languages, for example, provide technological resources that are often not provided by tech companies. Nonetheless, although many of these institutes are willing to share these resources, they frequently encounter legal barriers in the form of copyright and GDPR. Thus, quite a number of resources are unavailable to the private sector or the general public. In terms of needs, these institutions also note that there is neither enough access to data or AI and spoken dialogue systems, nor a sufficient number of trained professionals with expertise in both technology and a particular language. This issue is compounded by a lack of advanced LT for languages with fewer speakers and resources.

Openness is a recurring theme. Easier access to data and the construction of infrastructures involving all economic stakeholders are perceived as musts. Similarly, the development of open-source large language models for all EU languages is considered an essential starting point for reengineering traditional NLP tasks. This does not mean, however, that funding for research in rule-based, knowledge-based and hybrid approaches should be eliminated. In addition to promoting these goals, a future ELE programme should be open to

novel approaches, encourage experimentation, and anticipate emerging technologies. The projects selected for FSTP exemplify this line of thinking. The diverse results of these projects provided additional material, ideas, and expertise across a variety of domains, including more inclusive LT that incorporates sign language, lessons learned for the collection of multilingual audio data, a speech dataset for under-represented languages, the compilation and analysis of collected data concerning NLP/LT tasks and languages, a detailed list of computing facilities for LT, a BLARK for minority languages, and an artificial intelligence data kit. Such interdisciplinary research will be fundamental to the success of a future ELE programme as LT develops.

Finally, a prospective ELE programme should push for investment from industry alongside funding from governmental sources. And while member states must possess greater awareness about the need to invest more directly in LT for the languages within their borders, decisions concerning the appropriate extent of LT implementation for individual languages should reside in their respective language communities. One reason for this is that the latter are critical stakeholders when it comes to collecting data for languages with fewer speakers and each language requires specialised attention to address its particular needs. Experts in a determined language must decide which language resource categories should be prioritised, given that each language differs in terms of language resource distribution and a substantial number of languages possess no, or close to no, resources for some categories of language resources.

## References

- Itziar Aldabe, Aritz Farwell, and German Rigau. Deliverable D3.1 Report on new strategic documents and technology in the LT area and language-centric AI, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2023/01/ELE2\\_\\_Deliverable\\_D3\\_1.pdf](https://european-language-equality.eu/wp-content/uploads/2023/01/ELE2__Deliverable_D3_1.pdf). Project deliverable; EU project European Language Equality (ELE2); Grant Agreement no. LC-01884166 – 101075356 ELE2.
- Itziar Aldabe, Aritz Farwell, Athanasia Kolovou, Stelios Piperidis, Georg Rehm, and German Rigau. Deliverable 3.2 Missing resources and relevant stakeholders, April 2023. URL [https://european-language-equality.eu/wp-content/uploads/2023/05/ELE2\\_\\_Deliverable\\_D3\\_2.pdf](https://european-language-equality.eu/wp-content/uploads/2023/05/ELE2__Deliverable_D3_2.pdf). Project deliverable; EU project European Language Equality 2 (ELE2); Grant Agreement no. LC-01884166 – 101075356 ELE2.
- Oliver Blake. Deliverable D2.10 Report from LIBER, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_10\\_Report\\_from\\_LIBER\\_.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_10_Report_from_LIBER_.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.
- Maria Eskevich and Franciska de Jong. Deliverable D2.3 Report from CLARIN, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_3\\_Report\\_from\\_CLARIN\\_.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_3_Report_from_CLARIN_.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](http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf), 2018.
- Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way. Deliverable D1.3 Digital Language Equality (full specification), 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D1\\_3.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D1_3.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

- Federico Gaspari, Jane Dunne, Maja Popović, Andy Way, Maria Giagkou, Stelios Piperidis, Jana Hamrlová, Davyth Hicks, and Sabine Kirchmeier. Deliverable D1.3 Report on All Consultations with Stakeholders, 2023a. URL [https://european-language-equality.eu/wp-content/uploads/2023/06/ELE2\\_\\_Deliverable\\_D1\\_3.pdf](https://european-language-equality.eu/wp-content/uploads/2023/06/ELE2__Deliverable_D1_3.pdf). Project deliverable; EU project European Language Equality (ELE2); Grant Agreement no. LC-01884166 – 101075356 ELE2.
- Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way. Digital Language Equality: Definition, Metric, Dashboard. In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 39–73. Springer, Cham, Switzerland, June 2023b. In print.
- Guðrún Gísladóttir. Deliverable D2.7 Report from ECSPM, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_7\\_Report\\_from\\_ECSPM.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_7_Report_from_ECSPM.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.
- Jan Hajič, Tea Vojtěchová, and Maria Giagkou. Deliverable D2.5 Report from META-NET, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_5\\_Report\\_from\\_META\\_NET.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_5_Report_from_META_NET.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.
- Stefanie Hegele, Katrin Marheinecke, and Georg Rehm. Deliverable D2.6 Report from ELG, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_6\\_Report\\_from\\_ELG.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_6_Report_from_ELG.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.
- Maria Heuschkel. Deliverable D2.12 Report from Wikipedia, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_12\\_Report\\_from\\_Wikipedia.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_12_Report_from_Wikipedia.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.
- Halid Hrasnica. Deliverable D2.11 Report from NEM, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_11\\_Report\\_from\\_NEM.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_11_Report_from_NEM.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.
- Sabine Kirchmeier, Georg Rehm, Marie Mattson, Davyth Hicks, and Jane Dunne. Deliverable D1.2 Report on Consultations with Funding Agencies, Policy Makers etc., 2023. URL [https://european-language-equality.eu/wp-content/uploads/2023/05/ELE2\\_\\_Deliverable\\_D1\\_2.pdf](https://european-language-equality.eu/wp-content/uploads/2023/05/ELE2__Deliverable_D1_2.pdf). Project deliverable; EU project European Language Equality (ELE2); Grant Agreement no. LC-01884166 – 101075356 ELE2.
- Christopher Moseley. Atlas of the world's languages in danger, 3rd edn., 2010. URL Onlineversion:<http://www.unesco.org/culture/en/endangeredlanguages/atlas>.
- 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, June 2023. In print.
- Andrew Rufener and Philippe Wacker. Deliverable D2.4 Report from LT-innovate, 2022. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.
- STOA. Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March 2017. <http://www.europarl.europa.eu/stoa/>.

Marlies Thönnissen. Deliverable D2.2 Report from CLAIRE, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE\\_\\_Deliverable\\_D2\\_2\\_Report\\_from\\_CLAIRE\\_.pdf](https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D2_2_Report_from_CLAIRE_.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Andy Way, Georg Rehm, Jane Dunne, Jan Hajič, Teresa Lynn, Maria Giagkou, Natalia Resende, Tereza Vojtěchová, Stelios Piperidis, Andrejs Vasiljevs, Aivars Berzins, Gerhard Backfried, Marcin Skowron, Jose Manuel Gomez-Perez, Andres Garcia-Silva, Martin Kaltenböck, and Artem Revenko. Deliverable D2.17 Report on all external consultations and surveys, 2022. URL [https://european-language-equality.eu/wp-content/uploads/2022/04/ELE\\_\\_Deliverable\\_D2\\_17\\_Report\\_on\\_External\\_Consultations\\_-2.pdf](https://european-language-equality.eu/wp-content/uploads/2022/04/ELE__Deliverable_D2_17_Report_on_External_Consultations_-2.pdf). Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

## Appendix

### A Appendix: Full list of unique organisations that filled in the endorsement form

- (1) Ablio
- (2) ACD-Agency for Cultural Diplomacy
- (3) AFRD Georgia
- (4) Afûk
- (5) Association of Translation Companies
- (6) Athena Consultancy
- (7) Athena RC/ILSP
- (8) Audio-Visual Machine Perception Limited
- (9) Bangor University
- (10) Barcelona Supercomputing Center
- (11) BEE Inernational
- (12) BEIA
- (13) BeLazy Kft.
- (14) Brno University of Technology
- (15) Budapest University of Technology and Economics, Laboratory of Speech Acoustics
- (16) Cálamo&Cran
- (17) Centro de Inteligencia Digital, University of Alicante
- (18) Cambra d'Òc
- (19) Charles University
- (20) CiTIUS-Research Centre on Intelligent Technologies Univ. Santiago de Compostela

# Endorse the ELE Strategic Research, Innovation and Implementation Agenda

The overall vision and unifying goal of the ELE Programme is to achieve digital language equality in Europe by 2030. The ELE 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. Recently, the Strategic Research and Innovation Agenda (SRIA) and the ELE Programme were presented in a workshop in the European Parliament on 8 November 2022.

[Read the full ELE SRIA](#)

**We are primarily asking you for your endorsement of the ELE SRIA so that we can list you as a supporter on the ELE website.**

I support the ELE Strategic Research and Innovation Agenda.

Please provide some information about you and your organisation.

Figure 2: The SRIA endorsement form as published online (page 1/4)



I agree that my name and affiliation will be published on the European Language Equality website and that the data I submit through this form will be processed according to the [Privacy Policy](#). I understand that I can withdraw this consent at any time by informing the website operator (e.g. by e-mail or through the website contact form).

Your name (required)

Your email address (required)

Name of the organisation you represent (required)

Your main role in your organisation

Which of the following best describes the type of organisation you work for?

- Academic or research organisation
- Government (e.g., ministry, general secretariat etc.)
- Public body other than Government
- SME
- Large enterprise
- Non-Governmental Organisation (NGO) / Civil Society Organisation (CSO)
- Foundation / Charity
- Political Organisation
- Other

I want to receive the European Language Equality / European Language Technology newsletter.

If you have another five minutes to spare, we would be happy if you could assist us by giving us your views on the ELE recommendations!

Yes, let's do this.

Please name one or more of your organisation's initiatives, projects or programmes that contribute to achieving Digital Language Equality (please include URL, runtime etc.)

Figure 3: The SRIA endorsement form as published online (page 2/4)

**Please provide your comments on the ELE recommendations.**

**Policy Recommendations**

Reinforce European leadership and sovereignty in LT by establishing the ELE programme as a large-scale, long-term coordinated funding programme for research, development, innovation and education with the societal goal of digital language equality and the scientific goal of deep natural language understanding.

[Full list of Policy recommendations](#)

Comments:

**Governance Model**

Create a pan-European network of research centers to facilitate the coordination of the ELE programme at all levels and to strengthen awareness of the importance of lesser-used, regional and minority languages through various means and instruments, such as lobbying, the allocation of LT to the portfolio of an EU commissioner, the promotion of a distributed centre for linguistic diversity etc.

[Full list of Governance recommendations](#)

Comments:

**Technology and Data Recommendations**

Address the lack of available data by more systematic language data collection, by exploiting new methods for data generation, and by unleashing public language data while in parallel ensuring sufficient adaptations of technologies.

[Full list of Technology and Data recommendations](#)

Comments:

Figure 4: The SRIA endorsement form as published online (page 3/4)

**Infrastructure Recommendations**


Strengthen existing and create new research infrastructures and LT platforms that support research and development activities, including collaboration, knowledge sharing, and open access to data and technologies and to ensure access to GPU-based HPC facilities.


[Full list of Infrastructure Recommendations](#)

Comments:

**Research Recommendations**

Support research in Deep Language Understanding for a new generation of applications that can address the deeper questions of communication, common sense and reasoning. This research line should integrate speech, NLP, and contextual information as well as additional modes of perception and neurosymbolic approaches.



[About](#) ▾ [Strategic Agenda](#) ▾ [Open Call](#) [Deliverables](#) [Events](#) ▾ [News](#) ▾ [Contact](#) 

**Implementation Recommendations**

Initiate a 9-year coordination ELE programme, collaboratively funded by the EU/EC and participating countries, covering six LT themes and supporting each with coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs).

[Full list of Implementation recommendations](#)

Comments:

Figure 5: The SRIA endorsement form as published online (page 4/4)

- (21) CLiPS, University of Antwerp
- (22) CNRS
- (23) Cogniteva SAS
- (24) Cognuse Inc.
- (25) Croatian Language Technologies Society
- (26) Czech Language Institute of the Academy of Sciences of Czech Republic
- (27) Darmstadt University of Applied Sciences
- (28) De Taalsector
- (29) DEPARTMENT OF INFORMATION, University of Firenze
- (30) DFKI GmbH
- (31) Dictionnaire des francophones
- (32) Dublin City University
- (33) Dutch Language Institute
- (34) Emplocity SA
- (35) Estonian Language Council
- (36) Eurac research
- (37) EURECOM
- (38) European Federation of National Institutions for Language (EFNIL)
- (39) European Language Equality Network (ELEN)
- (40) Field 33
- (41) Fondazione Bruno Kessler
- (42) Freie Universität Berlin, University Library
- (43) Haaga-Helia University of Applied Sciences
- (44) HENSOLDT Analytics GmbH
- (45) HiTZ (UPV/EHU)
- (46) Hogeschool Utrecht
- (47) Horacio Saggion
- (48) Humboldt-Universität zu Berlin, Dept of Slavic and Hungarian Studies
- (49) Hungarian Research Centre for Linguistics
- (50) IBL, BAS
- (51) Impact Nation Institute
- (52) Incyta Multilanguage, SL

- (53) INDIRE
- (54) Institut d'Estudis occitans
- (55) Institut Grand-Ducal Luxembourg
- (56) Institute for Serbian Language SASA
- (57) Institute of the Lithuanian Language
- (58) Instituut voor de Nederlandse Taal
- (59) ISSA Polska
- (60) ITU Copenhagen
- (61) Jozef Stefan Institute, Ljubljana
- (62) Kamusi Project International and African Academy of Languages (ACALAN)
- (63) KantanAI
- (64) KTH (Royal Institute of Technology)
- (65) KU Leuven
- (66) L. Štúr Institute of Linguistics, Slovak Academy of Sciences
- (67) LangTec
- (68) Leiden University Medical Center
- (69) Lingsoft
- (70) Lingua Custodia
- (71) Logical Events Limited
- (72) Luxembourg Institute of Science and Technology
- (73) M47 Labs
- (74) MAMA AI Coolma
- (75) MAMA AI Eva, s.r.o.
- (76) Mastervoice
- (77) Member of the UK Parliament
- (78) Mid Sweden University
- (79) Miðeind ehf
- (80) Ministry of Culture
- (81) MIRSK
- (82) Mohamed Bin Zayed University of Artificial Intelligence
- (83) Moravská zemská knihovna v Brně
- (84) Mykolas Romeris University

- (85) National and Kapodistrian University of Athens
- (86) National Council for the Maltese Language
- (87) Network to Promote Linguistic Diversity
- (88) Neurolingo ([www.neurolingo.gr](http://www.neurolingo.gr), [www.neurolingo.com](http://www.neurolingo.com))
- (89) Ole Sol
- (90) Ollscoil na Gaillimhe
- (91) Pangeanic
- (92) Plataforma per la Llengua
- (93) Polish Academy of Sciences
- (94) Pontifical Catholic University of Peru
- (95) PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language
- (96) Riga Technical University
- (97) Romanian Academy
- (98) RWTH Aachen University
- (99) Secretary for Language Policy, Government of Catalonia
- (100) Semantic Web Company
- (101) Shangu Business Services sbs
- (102) Sirma AI (Ontotext)
- (103) Slator
- (104) Språkbanken (The Language Bank of Sweden)
- (105) State Language Department of Georgia
- (106) Tallers per la Llengua
- (107) Technical University of Darmstadt
- (108) Technical University of Moldova
- (109) Technological University Dublin
- (110) The Árni Magnússon Institute for Icelandic Studies
- (111) The Danish Language Council
- (112) TMServe
- (113) Trinity College Dublin
- (114) UCLouvain
- (115) UiT The Arctic University of Norway

- (116) UMONS
- (117) UNED
- (118) Universidad Carlos III de Madrid
- (119) Universidad de Jaén
- (120) Universidad de Zaragoza / Academia Aragonesa de la Lengua
- (121) Universidad del País Vasco / Euskal Herriko Unibertsitatea
- (122) Universidad Politécnica de Madrid
- (123) Universidade de Santiago de Compostela, CiTIUS Center for Intelligent Technologies
- (124) Universidade de Vigo
- (125) Università Guglielmo Marconi, Italy
- (126) Universitat de Barcelona
- (127) Universitat Oberta de Catalunya
- (128) Universität Osnabrück
- (129) Universitat Politècnica de València
- (130) Universitat Pompeu Fabra
- (131) University of Alicante
- (132) University of Applied Sciences in Konin, Poland
- (133) University of Applied Sciences Technikum Wien
- (134) University of Belgrade
- (135) University of Bologna
- (136) University of Copenhagen
- (137) University of Cyprus
- (138) University of Edinburgh
- (139) University of Galway
- (140) University of Groningen
- (141) University of Helsinki
- (142) University of Jaén (Spain)
- (143) University of Konstanz
- (144) University of Latvia
- (145) University of Malta
- (146) University of Sheffield
- (147) University of Surrey

- (148) University of Tartu
- (149) University of the Faroe Islands
- (150) University of Turin
- (151) University of Vienna
- (152) University of Washington
- (153) University of West Bohemia
- (154) University of York
- (155) University of Zagreb
- (156) University of Zurich
- (157) University Politehnica from Bucharest
- (158) Vicomtech
- (159) VSB - Technical University of Ostrava
- (160) Weber Consulting KG
- (161) Whispp
- (162) Wolters Kluwer Deutschland GmbH
- (163) ZLS (Zenter fir d'Lëtzebuerger Sprooch)

## **B Appendix: All feedback collected through the online endorsement form**

### **Policy recommendations**

- (1) To encourage mother-tongue teaching for speakers of official and non-official languages of the EU. To enable and empower European SMEs and startups to easily access and use LT in order to grow their businesses online independent of language barriers.*
- (2) Especially in light of security concerns, dependencies and sustainability it is imperative that Europe secures and extends its capabilities and resources concerning LT. Even more so in light of US and Chinese efforts.*
- (3) Greening all languages in digital space with ethical code2030. Preventing cyberbullying by enhancing transnational collaborations in RaD, multilingual AI 4 science diplomacy policy communication*
- (4) <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2023-human-01-03> is an excellent initiative.*
- (5) I fully support this because as a trained translator and language teacher in tertiary education, I am fully aware that everybody whose main professional focus is outside languages will find it very difficult to acquire communicative competences giving them equality with peers having English as heir first language. The inequality thus created is in sharp contrast to the democratic principles the European Union is based on.*



- (6) *I particularly like the development of BLARK-like minimum set of LT for a language. This is the fundamental building block for real equal participation of each language in the Digital Decade.*
- (7) *In addition to open access to resources and data, there should also be an initiative to prohibit misuse of data. This concerns both large-scale generative models used to create toxic and destructive use of language, as well as the right for minority and indigenous people's rights to their own data, and how it is used. Another aspect of this that is often missing is that the use of language technology isn't necessarily positive for a minority language community. By default, minority language communities are bi- or multilingual; in such a context it can be counterproductive, and even destructive to the language community to use LT tools to automatically produce machine translated text into that language. MT from that language to majority languages is a very different matter, and would normally be encouraged. The core point is that the consequences for the language community isn't always easy to see from the outside, and if necessary, the language community should have a way to say 'no' to certain uses of LT if they deem it detrimental to the health of their own language's future. This is especially true since LT performance and data availability is very different for minority and indigenous languages compared to the majority languages.*
- (8) *In times of AI, neural machine translation and the use of large language models in the language industry, it is key that distinction is made between human natural language and machine-produced output. Machine output can be true, untrue or partially true or untrue with correlations between data that is not always validated by humans. We add value to machine translation by providing full human post editing by native speakers and we add value in meetings by ensuring that the human authentic input is reliably and faithfully reproduced in the target language, the mission of human translators and interpreters. We support applications that combine the power of language technology with human validation in the process.*
- (9) *It should include actions to promote inclusion intended as removing comprehension and readability barriers for non-expert readers. Many readers are excluded from many texts today due to the level of complexity.*
- (10) *It's okay.*
- (11) *strong support of this recommendation!*
- (12) *The European Commission states that "the EU's partnership with Africa is a key priority for the Commission" in a multi-actor partnership guided by the EU and African Union (AU) Member States. In its Language Plan of Action for Africa, the AU clearly sets language equality as a pressing goal, and has identified digitalization as fundamental for its attainment. The needs and rationales for African language equality are exactly the same as those for European language equality, and the methods deployed to reach linguistic equity in Europe are largely similar to those proposed by Africa's designated intergovernmental language organ. Yet there has been no bridge between the abstract statement of "partnership" between Europe and Africa on the one hand, and on the other hand the concrete goal of language equality as implemented in ELE with respect to European languages and called for by the 55 AU Member States as equally important for the languages spoken by their citizens. It is incumbent upon the EU to recognize that the reasons for digital language equality in Europe mandate equal consideration in the realm of international cooperation, and to therefore extend support for the pursuit of language equality in Africa.*

- (13) *The lack of long-term directed funding has long played a central role in hampering speech and language technologies for under-resourced languages. Until now, minority and lesser spoken languages are often overlooked in terms of investment as their economic return isn't seen attractive enough. We cannot rely on such an economic and market driven approach to supporting languages and this is where improved government (national and EU) policy will be needed to make an impact. National governments must have the awareness and appreciation for investment into language technologies in order to include all sections of society and prevent further marginalisation.*
- (14) *The Policy Recommendations are very good. I would emphasize the need for the standardization of the BLARK set of language resources and capacities that all European languages should possess. The process have been started (UD initiatives) but it should continue covering all the basic language processing steps.*
- (15) *There are already partial programmes addressing LT, but those are nor large-scale neither long-term.*
- (16) *There is an increasing awareness of the dangers associated with LT. One is that large language models, like ChatGPT, are very prone to picking up human biases from the large volumes of text they are trained on. Another is the subjectivity inherent in any system for bias detection or fact checking. Another is that present low levels of NL understanding and common sense quite often result in catastrophic AI failures. Improved bias checking and deeper levels of NL understanding should mitigate some of these risks.*
- (17) *these sound more like goals than like recommendations.*
- (18) *To support the policy "LT as a free service"*
- (19) *We desperately need help from EU to reinforce our minority language.*

#### **Governance recommendations**

- (1) *Add the word "indigenous" to the description "lesser-used, regional and minority languages." The Sámi people of Northern Europe is the only indigenous people of Europe, and it seems apt to include that term in such descriptions. There are seven Sámi languages spoken within the borders of EU, nine if the Kola Peninsula of Russia is included. I suggest avoiding a too narrow focus on one technology (ie neural net, machine learning style LT). It limits both what languages can be covered, and the tools and services being considered. This is true for all sections.*
- (2) *Alocating the area of multilingualism, linguistic diversity and language technology to the portfolio of a EU Commissioner I find as very important political recognition of this field. It's a metaphorical foot in the doorstep for LT.*
- (3) *Although national languages are reasonably covered, there is a huge gap between "large" and "smaller" languages, and minority languages are often completely ignored (there are exceptions, though).*
- (4) *Any such activity should include idustry and industry labs (if you include industrial reserach labs in your statement of "research labs" abive that would be just fine. If you only consider academic ones, this falls short of what should be aimed for!*
- (5) *Art-based research, citizen science*

- (6) *Character coding in UTF-32, and ASCII based IPA symbols, like X-SAMPA, already cover most of world languages. It also has already been demonstrated that linguistic knowledge learnt from text in one majority language can often be transferred to allow learning from the smaller quantities of NL text available for minority languages. To avoid the need for separate keyboards for every language, work needs to be done to allow now not only character sets, but also displayed keyboard labels, to be dynamically switchable.*
- (7) *Evaluation and finance existing LT resources and incorporation of them in European Portals*
- (8) *Is ELE preparing a proposal for url <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2023-human-01-03>?*
- (9) *It's okay.*
- (10) *Major developments in AI, MT and MI (machine interpreting) will come from global bigtech companies focusing on English-centric applications for the consumer and the btb market. We believe that the inclusion of regional and less dominant languages in language technology will set the scene for creating a level playfield in economic, cultural and political participation, especially in light of the rapidly increasing influence of AI driven language technology.*
- (11) *The proposed Governance model is excellent with the brilliant idea of a multilingual LT benchmark, a European "SuperGLUE"-style shared benchmark, that tracks progress. Highly valued point is to create a pan-European network of research centers to facilitate the co-ordination of the ELE programme at all levels. Promote the idea of fair use of data to exonerate researchers of IPR violation threat, so that much more and clearer language resources should be built.*
- (12) *The web needs to reach out more directly to the speaking communities. That's where the corpus material lives. Among academics, linguists have strong contact with these communities because of years long relationships needed for data elicitation.*
- (13) *What do you mean with "research centers"? Are these dedicated centers for ELE research or will all stakeholders i.a. universities with focus areas in multilingual DH as well as university libraries aiming at "decolonising" anglo-centric knowledge infrastructures also be involved? Me and all colleagues who are lobbying for more language diversity in European knowledge infrastructures would appreciate the latter very much...*

#### **Data recommendations**

- (1) *Current open speech datasets are dominated by English/American language. This should be much ore diverse to be able to develop robust and accurate multi lingual models.*
- (2) *<https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2023-human-01-03> could help,....*
- (3) *I suggest adding something along the lines of the following: - ensure that all languages have access to basic language technology (keyboards/input methods, full character set coverage in all fonts, all languages being predefined on an OS level, proofing tools, etc) - ensure that all LT API's in operating systems and computing services are open, so that third parties can install and provide services for languages not covered by the OS or service producer - enforce a split between UI level "surface language" of software, and the underlying core software, such that the natural language part of software can be replaced or switched at runtime, and such that localisation of software can not be dictated by the software developer. It should not be the developer who determines what language a piece*

*of software should be available in, that should be a decision of the language community (possibly in cooperation with a localisation industry, see next). - fostering a solid industry for localisation of software independently of the software makers (see previous point)*

- (4) *Inclusion from data from different domains, cultural diversity-sensitive national language eco-system*
- (5) *Increasing data availability is of paramount importance. For languages outside of the top investment tier; sources such as corpora cannot provide adequate data, so methods to gather linguistic data directly from the speakers of a language should be pursued. Unfortunately, in the current environment, all attention is soaked up by AI, which is impossible for languages without significant datasets. For non-lucrative languages, it is essential to ignore the siren call of AI in the near term, and double down on collecting the bedrock data on which to build technologies for the future.*
- (6) *It's okay.*
- (7) *Making public sector data, data from broadcasters, social media, publishers etc. available to all is very important because these are valuable sources of multimodal data. Also, development of large open-source language models for all EU languages I find as the starting point for reinventing and reengineering all traditional NLP tasks and consequently for new tasks that are behind the current horizon. Setting the minimum language resources that all European languages should possess in order to prevent digital extinction is also crucial.*
- (8) *The EU can play an important role to establish guidelines and directives in applied language technology by involving economic stakeholders with a shared interest to promote, maintain and improve the universal values of language equality, diversity, inclusion and to avoid language inequalities and language barriers in the global economy.*
- (9) *The implementation of the Open Data Directive needs to be expanded to include language data as high value datasets, while ensuring that public bodies are informed of language data use and benefits. These changes will make data coordination at a national level in public administration a lot more efficient.*
- (10) *This is also key, but a special effort has to be made to make those resources fully available to a wide audience, and it means not using jargon, removing technical barriers and promoting accessibility via awareness-rising campaigns.*
- (11) *To enforce existence of open data updated catalogs with URLs and standard formats for data.*
- (12) *To facilitate the digitisation of large volumes of text documents from every language, improvements will be required in Optical Character Recognition accuracy and availability for different languages. The best available OCR tools, such as Google's Tesseract, need to be trained on more languages, and need to improve their recognition accuracy for small signs, such as punctuation and diacriticals, which are at present too often interpreted as noise and ignored. There is a rather important difference between 1.001 and 100.1*
- (13) *Very well covered this topic. Promote the idea of fair use of data to exonerate researchers of IPR violation threat, so that much more and clearer language resources should be built.*
- (14) *Web corpora already provide a lot of data for major languages, but minority languages with little web presence (there is a lot of them in Europe) are completely in the dark. This is especially visible in current Large Language Models that have no support at all for those minority languages.*

**Infrastructures recommendations**

- (1) *An open source infrastructure and ecosystem in applied language technology involving all economic stakeholders (industry, government, language service providers) is the best plan of approach to enable language equality in LT.*
- (2) *As mentioned above: it would be great if the creation of ELE research infrastructures would be organised as a participative process with stakeholders i.e. universities with focus areas in multilingual DH as well as university libraries aiming at "decolonising" anglo-centric knowledge infrastructures . Me and all colleagues who are lobbying for more language diversity in European knowledge infrastructures (DARIAH multilingual DH WG, national and local multilingual DH working groups) would appreciate a possibility for involvement very much...*
- (3) *Creation of an European network of centres of excellence in LT could be a R&D backbone for ELE.*
- (4) *Ensure the research centers funding for the access to the GPU-based HPC facilities is mandatory to unleash the creative end experimental power of academic research.*
- (5) *EsherCloud is willing to do that*
- (6) *GPUs are great, but don't forget the QPUs (quantum processor units) which are soon likely to be with us. It is often pointed out that quantum computers will not be best suited to every task in computation. However, they will be particularly well suited to many of the core computations which are already involved in machine learning, not least for training large language models.*
- (7) *HorizonEurope, Europe Cultural desk, Erasmus+*
- (8) *Mostly good, but as pointed out above, it focuses too narrowly on one LT paradigm. For languages with complex morphology and morphophonology (like the Sámi languages) in combination with next to no existing electronic text data (true for most minority and indigenous languages), there ought to be alternatives. We have proven in our work (see links at the top) that one can build solid LT tools using alternative technologies and LT paradigms. And just to be clear: even for the languages mentioned above machine learning has its place, especially when it comes to speech technology. It just can't be the only game in town.*
- (9) *Sustainable methods for language technology development need to be examined to reduce the need for so many HPC facilities.*
- (10) *To offer the processing and data resources for demanding NLP tasks for the development of free LT applications and resources*

**Research recommendations**

- (1) *Virtual corpora: Integration of corpora that will be accessible with common APIs. This will enable the access of big distributed corpora from a single point with one iteration mechanism. Chatbot collaboration to access and query open data. Moderator Chatbot that query and process the responses of different specialised chatbots.*
- (2) *All domains of the European project funds, national research centers, media*
- (3) *As above, too much focus on AI based LT. Machine translation: - develop hybrid or purely rule-based systems for languages with minimal data. - ensure interoperability between systems based on different technologies*

- (4) *Currently, state-of-the-art research (LLM) is done elsewhere and support for EU languages is almost a by-product of having enough (web) data. The lack of comparable EU-led research is very visible.*
- (5) *EU and EDF efforts could be better synchronized regarding these issues.*
- (6) *I don't find the Green LT as one of most important recommendations, but let it remain there among the others.*
- (7) *It's okay.*
- (8) *Large language models, human validated content, corpora, and open source language technology are required to offer a level playing field in applied language technology involving big tech language industry players.*
- (9) *Looks like <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2023-human-01-03> !*
- (10) *Specific lines of funding are required for speech and language technologies. These also need to highlight the importance of continued research for technologies for the minority languages of Europe.*
- (11) *The research should also address multilinguality and challenges of using non-Latin scripts in digital European knowledge infrastructures. → great chance to network & connect with research on disrupting digital monolingualism as represented in the following publications (selection): (1) Fiormonte, Domenico. 2021. "Taxation against Overrepresentation? The Consequences of Monolingualism for Digital Humanities." In *Alternative Historiographies of the Digital Humanities* edited by Dorothy Kim and Adeline Koh, 333–376. Santa Barbara: Punctum Books. (2) Paul Spence. (2021). *Disrupting Digital Monolingualism: A report on multilingualism in digital theory and practice (1.0)*. *Disrupting Digital Monolingualism (DDM)*. Language Acts & Worldmaking project. <https://doi.org/10.5281/zenodo.5743283>*
- (12) *These are ambitious and, some would say, still far away goals (not unlike nuclear fusion or fully autonomous vehicles). Any research program aiming for deep NLU and common sense, otherwise known as AGI (artificial general intelligence), should be carefully planned and preferably grounded in theory, so that its intermediate steps are reasonably likely to be achievable. Even with quantum computing and the internet, there are limits to the amount of data available and to the number of free parameters which any language model can practically use.*
- (13) *These recommendations are very good. I would add promoting the research into detected and removal of biased data and already built language models*

#### **Implementation recommendations**

- (1) *To have the EU establish binding legislation to liberate language API's and language interface strings etc from the developers of the underlying systems, so that language services (API's) and localisations can be provided for languages not served by the producers of the software systems. Feel free to contact me for further data, discussion and reasoning around this topic.*
- (2) *Don't let avatars fall between these 6 themes. Photo-realistic avatars will be widely used in language training and in communication, not least for deaf people. Also, present speech synthesis is all too often let down by the stress and coarticulation betraying an incorrect parsing of the sentence. It will therefore be necessary to combine deep NLU and context sensitive LLMs with speech synthesis.*

- (3) EU binding legislation to encourage or ensure participation of MSs as well as EU funding for multi-country projects is crucial.*
- (4) Industry involvement has to be guaranteed to make sense in the long-term, adoption and actual innovation in the field.*
- (5) Strongly support for the request for funding, but I do not know what exactly are the "six LT themes" "coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs)"? Is there a chance to involve the above mentioned researchers and knowledge infrastructure specialists (librarians, research software developers etc.) from the field of multilingual DH, decolonizing data in the Global North etc.?*
- (6) We are interested in participating in development programs for speech-to-speech applications with adaptive input from trained linguists. Ideally, this could lead to new business models for language service providers and practitioners (interpreters and translators) with post-editing or human validation.*
- (7) What about <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2023-human-01-03> ?*