



EUROPEAN LANGUAGE EQUALITY

D2.3

Report from CLARIN

Authors Maria Eskevich, Franciska de Jong

Dissemination level Public

Date 28-02-2022

About this document

| | |
|----------------------|--|
| Project | European Language Equality (ELE) |
| Grant agreement no. | LC-01641480 – 101018166 ELE |
| Coordinator | Prof. Dr. Andy Way (DCU) |
| Co-coordinator | Prof. Dr. Georg Rehm (DFKI) |
| Start date, duration | 01-01-2021, 18 months |
| Deliverable number | D2.3 |
| Deliverable title | Report from CLARIN |
| Type | Report |
| Number of pages | 37 |
| Status and version | Final |
| Dissemination level | Public |
| Date of delivery | Contractual: 28-02-2022 – Actual: 28-02-2022 |
| Work package | WP2: European Language Equality – The Future Situation in 2030 |
| Task | Task 2.1 The perspective of European LT developers (industry and research) |
| Authors | Maria Eskevich, Franciska de Jong |
| Reviewers | Maria Giagkou, Jan Hajič |
| EC project officers | Susan Fraser, Miklos Druskoczi |
| Contact | <p>European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland</p> <p>Prof. Dr. Andy Way – andy.way@adaptcentre.ie</p> <p>European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany</p> <p>Prof. Dr. Georg Rehm – georg.rehm@dfki.de</p> <p>http://www.european-language-equality.eu</p> <p>© 2022 ELE Consortium</p> |

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 | Athina-Erevnitiko Kentro Kainotomias Stis Plioroforias, Ton Epikoinonion Kai Tis Gnosis | ILSP | GR |
| 5 | Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country) | UPV/EHU | ES |
| 6 | CROSSLANG NV | CRSLNG | BE |
| 7 | European Federation of National Institutes for Language | EFNIL | LU |
| 8 | Réseau européen pour l'égalité des langues (European Language Equality Network) | ELEN | FR |
| 9 | European Civil Society Platform for Multilingualism | ECSPM | DK |
| 10 | CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium | CLARIN | NL |
| 11 | Universiteit Leiden (University of Leiden) | ULEI | NL |
| 12 | Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH) | ERSCM | DE |
| 13 | Stichting LIBER (Association of European Research Libraries) | LIBER | NL |
| 14 | Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e.V.) | WMD | DE |
| 15 | Tilde SIA | TILDE | LV |
| 16 | Evaluations and Language Resources Distribution Agency | ELDA | FR |
| 17 | Expert System Iberia SL | EXPSYS | ES |
| 18 | HENSOLDT Analytics GmbH | HENS | AT |
| 19 | Xcelerator Machine Translations Ltd. (KantanMT) | KNTN | IE |
| 20 | PANGEANIC-B. I. Europa SLU | PAN | ES |
| 21 | Semantic Web Company GmbH | SWC | AT |
| 22 | SIRMA AI EAD (Ontotext) | ONTO | BG |
| 23 | SAP SE | SAP | DE |
| 24 | Universität Wien (University of Vienna) | UVIE | AT |
| 25 | Universiteit Antwerpen (University of Antwerp) | UANTW | BE |
| 26 | Institute for Bulgarian Language “Prof. Lyubomir Andreychin” | IBL | BG |
| 27 | Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences) | FFZG | HR |
| 28 | København's Universitet (University of Copenhagen) | UCPH | DK |
| 29 | Tartu Ülikool (University of Tartu) | UTART | EE |
| 30 | Helsingin Yliopisto (University of Helsinki) | UHEL | FI |
| 31 | Centre National de la Recherche Scientifique | CNRS | FR |
| 32 | Nyelvtudományi Kutatóközpont (Research Institute for Linguistics) | NYTK | HU |
| 33 | Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies) | SAM | IS |
| 34 | Fondazione Bruno Kessler | FBK | IT |
| 35 | Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia) | IMCS | LV |
| 36 | Lietuvių Kalbos Institutas (Institute of the Lithuanian Language) | LKI | LT |
| 37 | Luxembourg Institute of Science and Technology | LIST | LU |
| 38 | Università ta Malta (University of Malta) | UM | MT |
| 39 | Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute) | INT | NL |
| 40 | Språkrådet (Language Council of Norway) | LCNOR | NO |
| 41 | Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences) | IPIPAN | PL |
| 42 | Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science) | FCULisbon | PT |
| 43 | Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy) | ICIA | RO |
| 44 | University of Cyprus, French and European Studies | UCY | CY |
| 45 | Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences) | JULS | SK |
| 46 | Institut Jožef Stefan (Jozef Stefan Institute) | JSI | SI |
| 47 | Centro Nacional de Supercomputación (Barcelona Supercomputing Center) | BSC | ES |
| 48 | Kungliga Tekniska högskolan (Royal Institute of Technology) | KTH | SE |
| 49 | Universität Zürich (University of Zurich) | UZH | CH |
| 50 | University of Sheffield | USFD | UK |
| 51 | Universidad de Vigo (University of Vigo) | UVIGO | ES |
| 52 | Bangor University | BNGR | UK |

Contents

| | |
|---|-----------|
| 1. Introduction | 1 |
| 1.1. About CLARIN | 1 |
| 1.2. Federated Service Offer, Alignment with the Open Science Agenda | 2 |
| 2. Methodology and Instruments for Collecting Input | 3 |
| 2.1. Online Survey | 3 |
| 2.2. Interviews | 4 |
| 3. Analysis of Responses to Survey Questions | 5 |
| 3.1. Respondents' Profiles | 5 |
| 3.2. Language Coverage | 6 |
| 3.3. Evaluation of Current Status | 8 |
| 3.4. Predictions and Visions for the Future | 9 |
| 4. Analysis of Interviews | 10 |
| 4.1. Evaluation of Current Situation | 10 |
| 4.1.1. Current Offer of Research Infrastructures in the Context of User Needs | 10 |
| 4.1.2. Multilinguality in Practice from the Point of View of Technology Development | 10 |
| 4.1.3. Development and Training of Digital Skills | 12 |
| 4.2. Predictions and Visions for the Future | 12 |
| 4.2.1. Speech and LTs for Users | 12 |
| 4.2.2. Immediate Challenges for LTs | 13 |
| 4.2.3. Legal and Administrative Support | 15 |
| 4.2.4. Human Resources and Initiatives for Education | 15 |
| 5. Conclusions | 15 |
| A. The LT Researchers and Developers Full Survey | 18 |
| B. Additional Tables and Graphs | 26 |

List of Figures

| | | |
|-----|---|----|
| 1. | Map of CLARIN members, observers, and participating centres (February 2022) | 2 |
| 2. | LT areas in which the respondents conduct research or develop tools and services | 5 |
| 3. | Type of organisation | 6 |
| 4. | Number of respondents that already work with the language and/or plan to process it in the upcoming three years | 7 |
| 5. | Full survey as published (page 1/9) | 18 |
| 6. | Full survey as published (page 2/9) | 19 |
| 7. | Full survey as published (page 3/9) | 20 |
| 8. | Full survey as published (page 4/9) | 21 |
| 9. | Full survey as published (page 5/9) | 22 |
| 10. | Full survey as published (page 6/9) | 23 |
| 11. | Full survey as published (page 7/9) | 24 |
| 12. | Full survey as published (page 8/9) | 24 |
| 13. | Full survey as published (page 9/9) | 25 |

List of Tables

| | | |
|----|--|----|
| 1. | Types of survey questions | 3 |
| 2. | Drivers for the decision to support additional languages that were mentioned amongst top three. | 8 |
| 3. | Breakdown of answers to “Which of the following best describes the type of organisation you work for?” (mandatory closed question) | 26 |
| 4. | Breakdown of answers to “Where is your organisation’s headquarter based?” (mandatory closed question, plus “if other” as optional open-ended question). The countries that are not CLARIN members are marked in Italics. | 27 |
| 5. | Breakdown of answers to “In which sectors are your technologies, products or services used?” (mandatory closed question, plus “if other” as optional open-ended question). | 28 |
| 6. | Breakdown of answers to questions Q14 and Q16 “What languages does your organisation conduct research in and/ or for what languages do you offer services, software, resources, models etc.?” and “Are there any languages that your organisation does not yet support, but you plan to support in the next three years?” respectively. | 29 |
| 7. | Answers to the question (Q20-Q28): “Please indicate if you agree with the following statements: “One of the main challenges and obstacles the European LT community currently faces is...” (mandatory closed question, answers provided on a four-point scale, plus “I don’t know/No answer”). The statements and numbers in bold represent the answers where the audience predominantly (more than 70 percent) agrees with the statement. | 30 |
| 8. | Answers to the question (Q30-Q38): “In your opinion, how effective can the following policies/instruments be in speeding up the development and deployment of LT in Europe equally for all languages?.” (mandatory closed question, answers provided on a five-point scale, plus “I don’t know/No answer”). | 31 |

List of Acronyms

| | |
|---------------|---|
| AI | Artificial Intelligence |
| AI4EU | AI4EU (EU project, 2019-2021) |
| CEF AT | Connecting Europe Facility, Automated Translation |
| CLAIRE | Confederation of Laboratories for AI Research in Europe |
| CLARIN | Common Language Resources and Technology Infrastructure |
| DARIAH-EU | Digital Research Infrastructure for the Arts and Humanities |
| DHCR | Digital Humanities Course Registry |
| DL | Deep Learning |
| DLE | Digital Language Equality |
| DH | Digital Humanities |
| EC | European Commission |
| ELE | European Language Equality (<i>this project</i>) |
| ELE Programme | European Language Equality Programme (<i>the long-term, large-scale funding programme specified by the ELE project</i>) |
| ELEXIS | European Lexicographic Infrastructure |
| ELG | European Language Grid (EU project, 2019-2022) |
| ELRA | European Language Resource Association |
| ELRC | European Language Resource Coordination |
| ELT | European Language Technology, communication channel for ELG and ELE |
| EOSC | European Open Science Cloud |
| ERIC | European Research Infrastructure Consortium |
| IoT | Internet of Things |
| LR | Language Resources/Resources |
| LT | Language Technology/Technologies |
| META-NET | EU Network of Excellence to foster META |
| ML | Machine Learning |
| MT | Machine Translation |
| NLP | Natural Language Processing |
| NLU | Natural Language Understanding |
| RI | Research Infrastructure |
| SRIA | Strategic Research and Innovation Agenda |
| SSH | Social Sciences and the Humanities |
| SSHOC | Social Sciences and the Humanities Open Cloud |
| VLO | Virtual Language Observatory |

Abstract

This report summarises the opinions and insights collected in the network, and highlights the important steps to be taken within the ELE Programme, and beyond 2030, in order to reach and to sustain digital language equality across Europe.

1. Introduction

This deliverable reports on the results and findings of a consultation with representatives of the Language Technologies (LT) community, i. e. industry and research/academia, conducted by the European Language Equality (ELE) project. The results documented in this report will serve as input for a strategic research and innovation agenda and roadmap, in order to tackle the striking imbalance between Europe's languages in terms of the support they receive through language technologies by 2030.

The ELE project collected the views of European researchers and developers to consolidate their perspectives regarding the strengths and weaknesses of the field and also regarding the measures that need to be taken, so that all European languages are equally supported through technology by 2030. This diverse group of stakeholders comprises:

- Academic and industrial researchers in the field of LT/NLP – beyond pure research, they develop algorithms, pre-commercial LT prototypes, applications and systems
- Innovators and entrepreneurs who commercialise LT to address the needs of digital content analysis and generation, pertinent content transformation and dissemination, as well as enhanced human-machine interaction.

The field of Language Technology stands at the intersection of Linguistics and Computational Linguistics, Computer Science and Artificial Intelligence, while at the same time it encompasses methods and findings from Cognitive Science and Psychology, Mathematics, Statistics, Philosophy and more. Due to this **multi- and interdisciplinary** nature, the ELE stakeholders group of LT developers also includes neighbouring disciplines, especially AI and Digital Humanities/Social Sciences and Humanities (DH/SSH). To reach out to this diverse and extensive group of stakeholders, the ELE consortium invited representatives of various European networks, associations, initiatives and projects covering both research and industry to participate in a survey. In addition, input on specific topics was collected through a written interview to a subset of the survey respondents. A common methodology and set of instruments was utilised to carry out the survey, analyse its output and conduct the interviews across all communities approached. This report covers and analyses responses and input from members of the (CLARIN) infrastructure.¹

1.1. About CLARIN

CLARIN (Common Language Resources and technology Infrastructure) is one of the pan-European research infrastructures (RI) that form the RI landscape that is supported and monitored by ESFRI.² It is strongly rooted in the humanities and the field of Natural Language Processing (NLP) and has the mission to create and maintain an infrastructure to support the sharing, use and sustainable availability of language data and tools for research in the

¹ Reports from other groups of ELE stakeholders will be published on the ELE website (<https://european-language-equality.eu>), as they become available.

² <https://www.esfri.eu/esfri-roadmap-2021>

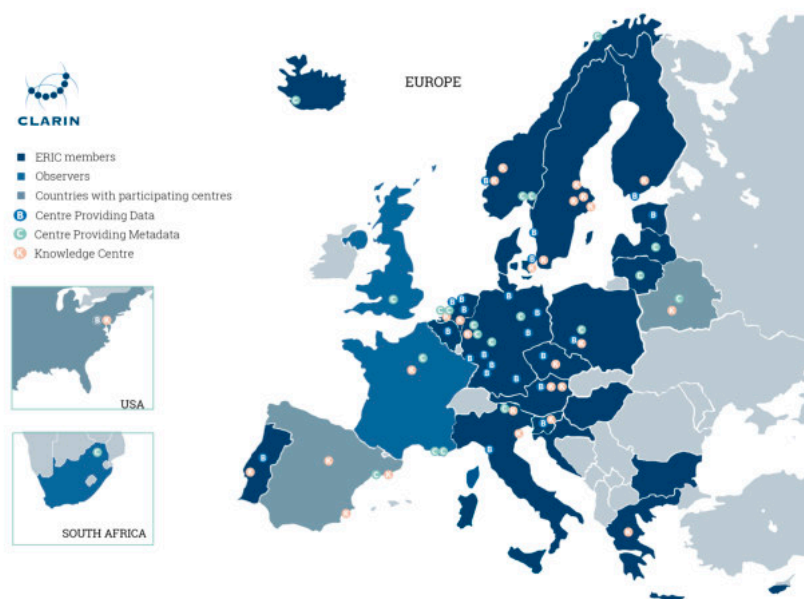


Figure 1: Map of CLARIN members, observers, and participating centres (February 2022)

Social Sciences and Humanities (SSH) and beyond.³ Since its early days, the CLARIN consortium has aimed at building both a technical infrastructure and a sustainable organisation for collaboration and coordination across the participating national consortia, as well as the exchange of knowledge and best practices, see Broeder et al. (2008); Hinrichs and Krauwer (2014). The CLARIN infrastructure is adhering to the interoperability paradigm on several levels, including metadata harmonisation and standardisation, see de Jong et al. (2020).

The CLARIN consortium was established as a legal entity in 2012. It is a so-called ERIC (European Research Infrastructure Consortium), which is based on a model for funding and governance by the participating parties, with room for in-kind contributions from national consortia and independent third parties, both from Europe and beyond. Figure 1 shows the geographical spread of all CLARIN member countries (21), observer countries (3), third party (1) and distributed network of centres.

1.2. Federated Service Offer, Alignment with the Open Science Agenda

The access to language data and tools provided by the CLARIN infrastructure is organised through the model of service federation based on a distributed network of centres. In the wider landscape of disciplinary (otherwise known as thematic) infrastructures, the CLARIN service offer is aligned with the developments of the European Open Science Cloud (EOSC)⁴ with the ambition to contribute to the acceleration of the wider accessibility and reusability of data for research purposes and the wider Open Science agenda through a strong focus

³ See <https://www.clarin.eu/content/vision-and-strategy>

⁴ <https://www.clarin.eu/eosc>

on making data FAIR.⁵ The thematic language processing services offered by CLARIN centres can be applied to the thousands of digitised language resources that are accessible online. The CLARIN infrastructure helps to discover the resources available in the network of more than twenty-five certified data centres and serves as matchmaker for the data sets and tools. The collaboration with several other RIs in the European landscape is partly aimed at developing a strong basis for supporting multidisciplinary research agendas for all fields in which language is a relevant data type. Alongside the CLARIN Virtual Language Observatory (VLO),⁶ in particular the SSH Open Marketplace,⁷ developed in the H2020 project SSHOC,⁸ will become an important discovery platform for language materials. Reaching digital language equality (DLE) would enable the RIs in the European landscape to reach out to even more disciplinary communities and will thus increase the thematic service providers' potential impact.

2. Methodology and Instruments for Collecting Input

2.1. Online Survey

The survey was addressed to LT researchers and developers and aimed to elicit their views and insights on the state of digital language equality. The survey structure was geared towards the analysis, consolidation and integration of the collected feedback into the ELE SRIA and roadmap.

It encompassed forty-five questions in total, some of which depended on previous answers. As a result, respondents were presented with thirty-two (minimum) to forty-five (maximum) questions, including the 'if other' questions. Thirty-five questions were mandatory and twenty-seven were closed questions (single or multiple choice; see Table 1).

| | Mandatory | Optional | Total |
|------------|-----------|----------|-------|
| Closed | 24 | 3 | 27 |
| Open-ended | 2 | 16 | 18 |
| Total | 26 | 19 | 45 |

Table 1: Types of survey questions

The survey was structured in four main parts:

- **Part A. Respondents' profile:** The first part of the survey included thirteen questions for the demographic profiling of respondent, with emphasis on characteristics relevant to the task at hand, i. e.
 - Country
 - Affiliation
 - Type of organisation
 - LT areas that the respondent is mainly active in
 - Participation/membership in networks/associations

⁵ FAIR is short for: Findable, Accessible, Interoperable, Reusable; see de Jong et al. (2018) and Wilkinson et al. (2016) for examples and background.

⁶ www.clarin.eu/vlo

⁷ <https://marketplace.sshopencloud.eu>

⁸ <https://www.sshopencloud.eu>

- Sectors/domains that the respondent is active in (if relevant).
- **Part B. Language coverage:** The second part investigated the degree of coverage of the European languages in the context of the respondents' current research and development activities, i. e.
 - Languages currently supported in research/products/services
 - Languages planned to be supported in the short-/middle-term
 - Factors that influence the respondents' choices and decisions with regard to language coverage and support development.
- **Part C. Evaluation of current situation:** This part included questions that sought to elicit the respondents' evaluation of the status of LT research and development, the strengths, the gaps and the challenges that the European LT community is facing, i. e.
 - Gaps in terms of technologies, tools or applications, and resources, especially with regard to specific languages
 - LT areas where the European LT community excels
 - Main perceived challenges and obstacles that should be overcome.
- **Part D. Predictions and visions for the future:** The fourth part of the survey is the forward-looking section that investigated ideas, predictions and wishes of the LT community about how the LT field as a whole will achieve equal support for all European languages by 2030, i. e.
 - Policies/instruments that could help to speed up the effective deployment of LT in Europe equally for all languages
 - Prediction of future opportunities for LT in basic and applied research (scientific vision) and in innovation and the industry
 - Expectations of the community with regard to the challenges an ELE programme can address by 2030.
- **Follow-up:** The last three questions asked the respondents' permission to be contacted for an interview and, given an affirmative answer, their contact details.

The survey was designed within the ELE project, and set up and published on the EU Survey platform.⁹ The full survey, as published online, is presented in Appendix A (p. 18 ff.).

To collect the input for this report, the survey was distributed through emails to all CLARIN national consortia and centres, and the overall importance to express their opinions in detail was explicitly emphasised. It was also advertised through the ELE, ELG and ELT websites.¹⁰

The survey was opened on 17 June 2021 and closed on 18 October 2021. In total, 333 responses were collected. Out of 333, ninety respondents indicated that they answered the questions in their capacity as member of the CLARIN network. This particular subset of responses is analysed in this report.

2.2. Interviews

Four interviews were conducted via email in the period between November 2021 and January 2022. The informants were selected to represent four different angles when answering the questions of the value of achieving DLE:

⁹ <https://ec.europa.eu/eusurvey/runner/ELE-LTdevs>

¹⁰ <https://european-language-equality.eu>, <https://www.european-language-grid.eu>, <https://www.european-language-technology.eu> as well as through the ELT social media accounts on Twitter and LinkedIn.

- Perspective of users of Research Infrastructures such as scholars from Social Sciences and Humanities (SSH) at large and beyond
- Multilinguality in practice from the point of view of technology development
- Translation studies, data and tools
- Development and training of relevant digital skills.

The informants were asked to provide more extensive responses for parts C and D of the original survey, and they were informed about the angle they should focus on. Their answers were synthesised retaining as much of the original text as possible.

3. Analysis of Responses to Survey Questions

3.1. Respondents' Profiles

At least one representative from each national CLARIN consortium filled in the survey. Input was also received from Switzerland and Luxembourg. The number of respondents per country ranged from one to ten, e.g. Czechia and Greece are represented by ten and nine respondents respectively.

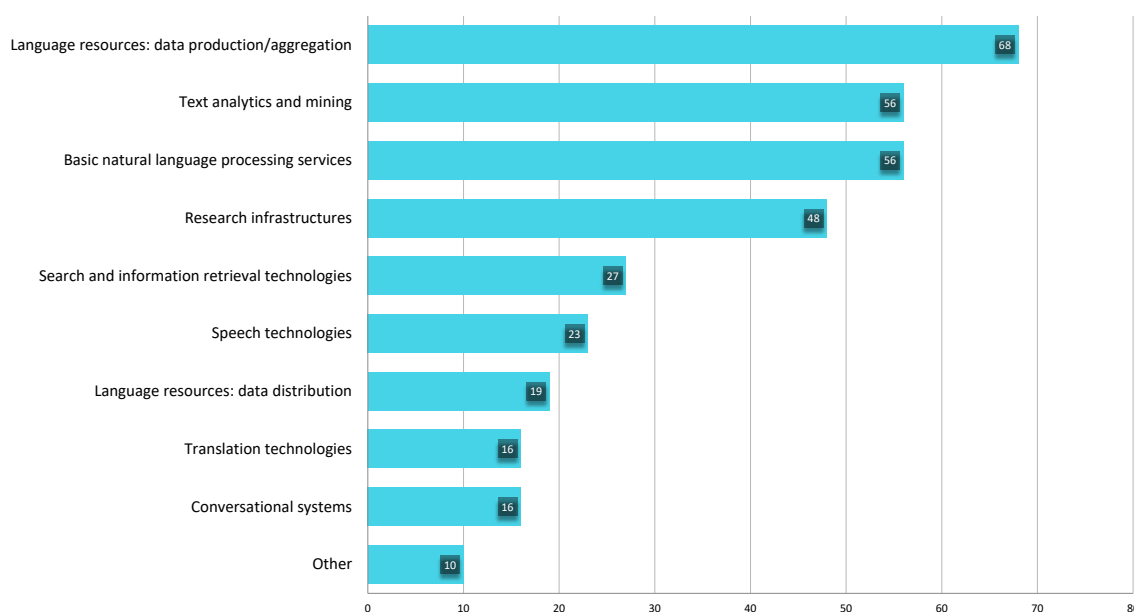


Figure 2: LT areas in which the respondents conduct research or develop tools and services

Out of ninety respondents, two are from an SME, while others represent the diversity of the research/academia sectors, varying from universities to research centres, libraries, and a memory institution (seventy-seven different organisations in total) (Figure 3). The SMEs are from Latvia (TILDE) and Finland (LingSoft). Detailed statistics of the breakdown of organisation types and countries are provided in Appendix B (Tables 3 and 4).

The respondents are mainly active in the following LT areas:

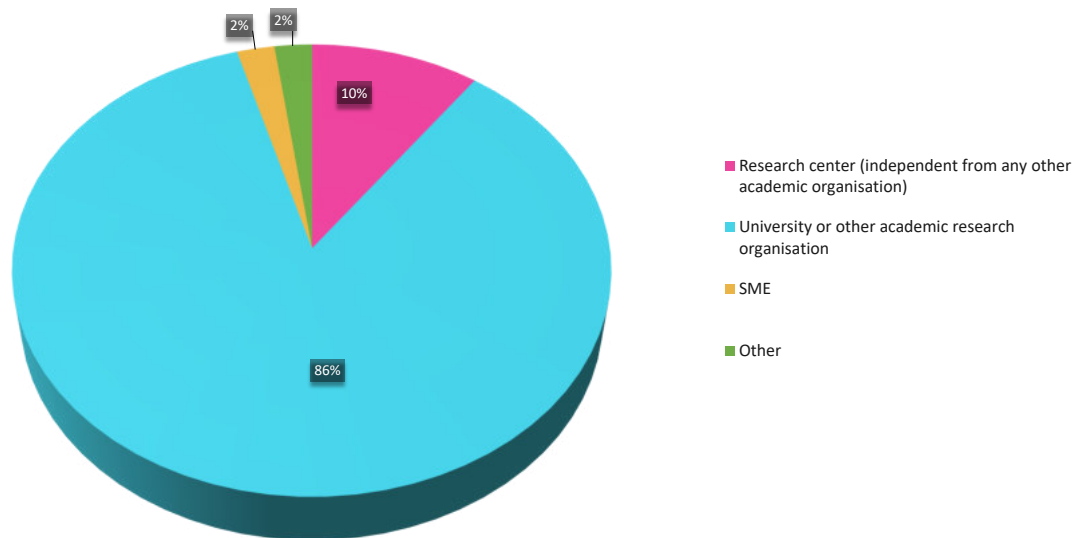


Figure 3: Type of organisation

- Language resources (data production and aggregation, as well as data distribution)
- Text analytics, mining, information extraction
- Basic NLP services.

Figure 2 reflects the variety of CLARIN services on offer, as the respondents work with different modalities (text and speech), different languages (including studies in machine translation), diverse aspects of data analytics (from text and data mining (TDM) to information retrieval).

The broad scope of the CLARIN service offer is demonstrated by the variety of application domains that the respondents have indicated. The most frequently mentioned domains include: digital humanities, arts, culture and other services; education; information and communication technologies; social sciences; and health.

For exact numbers, see Appendix B, Table 5.

3.2. Language Coverage

CLARIN serves a large multilingual community and the respondents reported that they already work with the thirty-three official European languages that were listed in the survey. Moreover, their research interests already span more than ninety-eight other languages, representing various historical and cultural perspectives, such as Early Modern Swedish, Early New High German and church Slavic. They also include a broad geographical spread, covering the languages of all continents. Furthermore, there are plans to start working with at least eight more new languages in the upcoming three years. See Figure 4 for a general overview and Appendix B, Table 6 for the full list of languages.

Three dominant drivers to select certain languages to work with include research and scientific interests, the availability of funding/investment, and the availability of language resources. The CLARIN infrastructure and service offer is primarily focusing on providing support for the scientific community. This is in line with the finding that market interests have

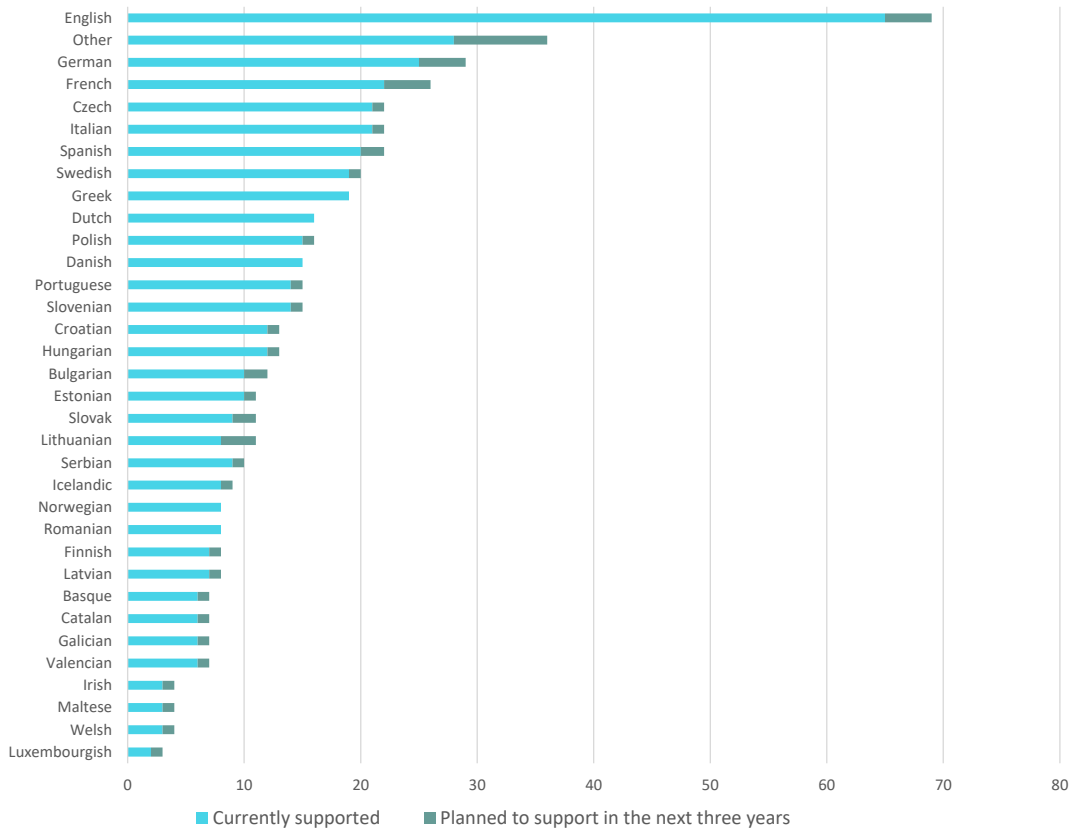


Figure 4: Number of respondents that already work with the language and/or plan to process it in the upcoming three years

a secondary role in the responses. Two respondents indicated to be motivated by two additional drivers: language preservation, and language policy and cultivation considerations. It is also noted that certain institutions were founded to conduct research in one particular national language, and therefore there is a legal restriction not to broaden the scope further to incorporate other languages. The exact numbers across the respondents can be seen in Table 2.

| Drivers | Answers |
|---|---------|
| Research/scientific interest | 77 |
| Available funding/investment | 47 |
| Availability of language resources | 42 |
| Availability of human experts for other languages | 26 |
| Market interest/demand by users or customers | 22 |
| Availability of technologies/tools | 18 |

Table 2: Drivers for the decision to support additional languages that were mentioned amongst top three.

3.3. Evaluation of Current Status

The majority of the CLARIN network respondents see two types of issues that represent the main challenges and obstacles for the broader European LT community. On the one hand, the focus on the development of new technologies shifts attention from the fundamental research, even though this type of research can provide a solid base for future technology development. While in recent years some advanced technology opened by mostly non-European commercial players such as Google, Microsoft, Amazon and Facebook propelled LT research, and fuelled advancements across the NLP and AI fields, the next steps of more complex tasks cannot rely only on a ‘throw more data at it’ approach. Such extensive analysis can be rooted in FAIRified research. On the other hand, the importance of multilinguality on the European landscape does not always get adequate recognition, and the smaller languages appear not to be attractive enough to receive the level of attention that is required to reach DLE in the long term. Very often one can rely on public funding only (practically) to work on the smaller languages, as there is no market for such LTs. These public investments for small languages are necessary on a larger scale to really make them available to the wider community.

The cost of developing LT for a language is usually constant, regardless of the number of speakers of that language. Even more, for languages with larger numbers of speakers, the LRs could be collected in an easier manner: for instance, the larger the number of speakers, the more online-collectable text is produced in a day. Industry can find a commercial interest in pre-competitive investments for ‘larger’ languages, while this will rarely be the case for ‘smaller’ ones. In that situation, the role of additional investor for the development of LT for ‘smaller’ languages should be played by bodies either at national or EU level. This situation is even worse for non-standard languages: local dialects, non-standard written language on social media platforms, non-standard language for speech recognition, and non-standard language as used by migrants or citizens with a migration background. There are almost no calls to fund work on creating language resources for training models or the research into these languages and modern approaches, such as deep learning, which require large computing capacities, would also often lie outside of the financial scope of smaller research institutions.

As CLARIN network members work across most of the European countries, their impression regarding the talent drain is far from homogeneous, as some of the countries might benefit from talent mobility more than others. The talent drain can be also observed between scientific fields. LT is seen as a subset of ML tasks, and the high demand for ML specialists on the market deflects potential talents from the LT field, and influences both the planning of the curriculum, and students' selection of the field to specialise in.

The respondents mention copyright protection that strongly reduces access to language data. Different governments appear to have transposed and implemented the GDPR in relation to research differently, and it would be beneficial for the EU's ability to advance language research if it was clearer on how the regulations should be interpreted nationally in a consistent way in terms of research.

A detailed list and more exhaustive summary of all answers can be found in Appendix B, Table 7.

3.4. Predictions and Visions for the Future

Over the course of more than a decade of CLARIN existence, the continuous investments in a research infrastructure that supports LT have proved to be a very efficient way to ensure the continuous research workflows, and to reinforce the application of novel, data-driven methods in broader domains such as SSH. This, in turn, stimulates the LT development for the use cases that are beneficial to researchers at large, thus increasing the societal impact. Moreover, implementation of LTs for diverse communities and different stakeholders increases the user base of multilingual systems.

Long-term support for infrastructural initiatives allows to build and efficiently implement strategies to reach interoperability between data and services. As introduced in Section 1.2, the work on EOSC development provides the foundation for FAIR research in the future.

An important direction for the future development of LTs that could lead to DLE lies in the domain of education, training and in the core of the scientific system. To start with, explanations of the difference and connection between LTs, ML, and AI should be given to diverse audiences, from school level to popular science. These topics should be part of the curriculum at different levels. This knowledge sharing should be supported through the actual use of LTs across EU and national organisations, as such common practice would solidify the understanding of their value. Researchers working on the non-English LTs should have appropriate venues to publish their research, and open access journals should be recognised and rewarded in national research assessments to boost the growth and appreciation of talents in the field.

A detailed list and more exhaustive summary of all answers can be found in Appendix B, Table 8.

Looking ahead, the community sees the following directions that require more time and effort in order to reach DLE:

- **Harmonisation and adoption of standards** is at the core of RIs work. Collaboration through cluster initiatives within domains (such as SSH), and further under the umbrella of EOSC builds ground for smooth integration and wider adoption of standards
- **Explainability of LTs** that leads to its higher trustworthiness. Recently many instances showcase the impact of training data bias on the systems output
- While stating the vision for LTs one should not shift focus from the **human experts** who stand behind those, and require adequate funding schemes for both advanced development and support work

- **Legal and administrative support** for the researchers who need to access, work with, and further share different type of content
- **Green LT** (i. e. technologies with low-demand computational footprint). Access to the data and tools via distributed RIs is the starting point to optimise the footprint.

4. Analysis of Interviews

In order to elaborate on the views and opinions in the network, four colleagues from the network have been chosen for a more detailed discussion on their understanding of the current situation, and on their vision into the future. The answers of the interviewees are grouped together, yet keep as much of their insights as when they were originally expressed.

4.1. Evaluation of Current Situation

4.1.1. Current Offer of Research Infrastructures in the Context of User Needs

As the initiative *CLARIN Resource Families*¹¹ shows, the availability of the basic resources, models and tools is quite satisfactory for the processing of written standard language for most official European languages, that is, for most of the official languages at national and European levels. However, the resources for spoken as well as all non-standard varieties, which are becoming increasingly important due to the widespread mass communication platforms and technologies, are virtually non-existent. In order to ensure an equal playing field for European citizens in the increasingly digital world, these need to be made available and accessible to academia as well as the industry sector as soon as possible. What is more important, coordinated efforts with respect to data collection, annotation and encoding are required, in order to be able to maximise the training and reuse potential of the tools and services for multiple languages, as well as enable cross-lingual and transnational research. These efforts can be enabled and coordinated through the established research infrastructures for sharing LRs, such as CLARIN.

4.1.2. Multilinguality in Practice from the Point of View of Technology Development

Currently, the status of LT technologies development can be classified along several dimensions: type of resource, tool or technology; the domain, that is, whether it is general or specialised, such as healthcare, automotive, digital humanities, or social interaction; multilinguality, that is, for which languages the support is strong or weak, or even nonexistent; quality, which is typically related to data availability for the tasks, domain or language.

In general, one of the main issues on the path to reach DLE is the usage of the concept of 'size' to define the importance of the work with the language, and whether the current volume of resources and tools is sufficient or not. This 'size' could be ascertained in different ways, for instance based on the overall number of speakers, or the GDP of the countries where the language is a major language. There are some exceptions, but overall the correlation is present. The smaller countries have (obviously) less universities and research teams to do such research and to develop resources and technologies for their languages, thus further widening the gap for those languages. The same situation holds for applications, regardless of whether or not they are LT-heavy, such as machine translation, or simply using a language component, such as voice commands in cars. The decisions by the providers are based on business considerations, and the 'large' languages have priority. We do see more

¹¹ <https://www.clarin.eu/resource-families>

and more smaller languages covered, but the process is slow and the quality of the applications for the smaller languages is lower, sometimes even substantially.

In the Social Sciences and Humanities, with a wide range of tasks, resources, methods and approaches to research, the situation is equally uneven. This holds true not only for linguistics, in which language per se is studied, but also for other disciplines in which language data is studied with respect to the information it conveys. Diachronic language changes can pose major problems when using language models trained on the available contemporary data which might differ in terms of script, orthography, morphology, lexicon and even syntax, and thus be less reliable and useful in terms of produced output. Limited online availability of language data sets required for research purposes is another problem. For example, subscription magazines and books might be not available for potential model training. The situation is even worse when it comes to annotated data, even if only simple annotation is considered, such as part-of-speech for most historical languages, with the exception of Latin and Old Greek, which are studied broadly and for which annotated data do exist. It is also very difficult to find transcription of historical speech, for example for radio broadcasts from before the period when archival efforts became widespread (in the 1950s and 1960s). These limitations all negatively affect not only language studies, but all research domains in which the availability of digital content in textual or spoken format in multiple language is a necessary precondition for the enabling of comparative research.

Across the LTs, the situation usually appears to be better in areas where enough data is available, which is primarily the case for machine translation. MT systems benefit from the fact that EU institutions translate and publish their documents in many EU official languages, though the number of available EU resources for those languages still depends on how long ago the country was admitted to the EU. Today, eTranslation is in use by 108 projects – eighty-seven projects reusing eTranslation and twenty-one projects committed to analysing or re-using eTranslation. For example, European translation technologies support translation needs during the EU Council presidency in many countries through the EU Council Presidency Translator (Pinnis et al., 2020) and so do public administrations (e.g. public administrations in Finland, Estonia and Latvia through the NTEU project). Different translation services are available from the CLARIN VLO and corresponding national nodes (e.g. from LINDAT/CLARIAH-CZ (Kořarko et al., 2019), PORTULAN CLARIN, and CLARIN-IS (Snæbjarnarson et al., 2021)). In other words, machine translation – even if not perfect – is now available for all official EU languages. For the 'big' languages (English, French, German, Italian, Spanish, often also Portuguese), many other resources, and therefore basic tools, are also available – speech recognisers, basic language resource toolkits (POS taggers, syntactic analyzers), and even more application-oriented tools such as sentiment analysis, named entity recognition, entity linking, information extraction etc.). Gaps still exist in terms of small languages (except for the basic language resource toolkits that are now available for 120+ languages thanks to the Universal Dependencies databases) and application areas, where business needs still prevail. Limited MT support by the European LT industry for specific language pairs and domains is a reason why European citizens in many cases still rely on global companies and external providers outside of Europe (Vasiljevs et al., 2019).

Overall, current challenges in translation technology include language equality (e.g. less resourced languages and domains are weakly represented), deep natural language understanding, simultaneous translation/interpretation of the spoken language, data availability (lack of training data for specific language pairs or domain) and technological requirements for huge volumes of data, as well as limited infrastructural resources (e.g. GPU and TPU).¹²

¹² Multilinguality and Machine Translation are elaborated in the ELE deliverable *D2.13 Technology Deep Dive – Machine Translation*, available at <https://european-language-equality.eu/deliverables/>.

4.1.3. Development and Training of Digital Skills

Event though education itself is not a core activity of RIs, CLARIN supports development and training of digital skills for different levels of researchers through a number of initiatives.

- CLARIN has started a ‘Teaching with CLARIN’ platform¹³ which encourages the integration of CLARIN resources, tools and services into the curricula of SSH-related disciplines. It allows and encourages teachers and lecturers to publish training materials in open access, to share best practices in teaching, and to reach out to those who would like to reuse the work prepared by their colleagues.
- Through collaboration activities between RIs, CLARIN ERIC and DARIAH-EU enable knowledge exchange and researcher mobility in the domain of DH. The Digital Humanities Course Registry (DHCR)¹⁴ is an example of such collaboration. This platform contains a selection of DH courses offered by European academic organisations, and allows students, lecturers and researchers to search on the basis of disciplines, topographical information (location), ECTS credits or the academic degrees that are awarded.

While the trend to have growing numbers of university-level educated researchers and software engineers in the fields of ML and DL has been stable in the past decades, AI and/or NLP courses oriented towards ML are now being taught beyond technical universities to broader audiences. This encourages more diversity of expertise and research questions in the field. However, the hype around AI and the overwhelming interest in the topic has not prevented a strong ‘brain drain’ to US universities, and in some cases, also to Asian ones, for experts with training at PhD-level or higher.

4.2. Predictions and Visions for the Future

Predictions for the future rely on three pillars: (i) scientific development that provides a pathway towards the support of the research questions and LTs beyond the year 2030 towards which the ELE Programme outlines the actions (Sections 4.2.1 and 4.2.2); (ii) legal and administrative support that are arranged at EU level and implemented at the level of RIs (Section 4.2.3); and (iii) attention for adequate capacity in terms of human resources (Section 4.2.4).

4.2.1. Speech and LTs for Users

As speech is one of the easiest and most natural ways to communicate, the main challenge and area which should be in focus in the near future for the European LT community is the shift from text to speech processing. This will have a direct impact on the lives of European citizens and will improve a broad range of public as well as business services in an unprecedented way, and will also improve access and ensure equal participation to citizens with special needs, such as the elderly, hard of hearing, or visually impaired. However, this can only be achieved if the resources, language models and technologies are developed on the foundations of equality, fairness and ethics.

The broadest opportunity lies in adding LT (especially including speech) technology to any product where human-computer (human-device, human-system) communication is the natural way of transferring information, providing feedback, asking questions, and so on. In other words, we will be seeing a shift from ‘LT-heavy’ applications (such as machine translation, except perhaps for speech translation, which still has not made it into the mainstream) to smaller ‘modules’ that will enhance everyday communication and control in various areas

¹³ <https://www.clarin.eu/content/teaching-clarin>

¹⁴ <https://www.clarin.eu/content/dh-course-registry>

– from cars to businesses and smart homes, social services, health care, public administration, etc. It is equally important to make these tools suited to safety and security, and to make them available – with the usual caveats for privacy – to defence and internal security institutions.

Already today translation technologies are widely used in general and by language specialists and language service providers. The use of translation technology will definitely grow, covering new application areas (e.g. IoT, smart homes and other smart devices), markets, supporting the Digital Single Market and language equality. The future translation technologies need to be able to dynamically adapt to the situation and context, and to be able to use general and culture-specific knowledge. Two main dimensions for opportunities could be domains (such as social media data translation, translation in emerging situations between complex, less/low resourced languages) and modalities (interpretation, speech translation, live subtitling and translation).

4.2.2. Immediate Challenges for LTs

Some of the challenges stem from previous observations. In the current technology environment, it is the data that matters most. Thus, one of the main tasks is to make sure that sufficient amount of data is available in sufficient quality, clearly licensed, easily accessible and usable from research to industry (including small companies), for all the tasks that are of relevance to both research and industry. Equally important is that European companies can grow and become successful (beyond being sold to large multinational ones), as e.g. SAP has demonstrated in the past. There are numerous startups in AI and Language Technology, but virtually none have grown to reach global importance.

There are several areas where public intervention might help:

- Funding data collection, annotation and distribution efforts, mainly through established research infrastructures and/or established ‘marketplaces’ where data can be found, accessed, downloaded, and used by both research and industry with clear licensing conditions.
- Further and expanded funding for both fundamental and experimental, cutting-edge research, that would keep a highly qualified, excellent research and teaching workforce in Europe to educate the next generation of researchers in AI and LT, and which would allow to engage PhD students as junior researchers on these projects, to get hands-on experience with state-of-the-art methodology, datasets and algorithms, so that they are prepared for the research and application challenges lying ahead.
- Provision of intervention instruments for companies which proved successful as startups and have a chance to grow substantially, by supporting innovation hubs, networks of investors, networks of capable managers willing to step in for executive roles, and by supporting them through ‘buy Europe’ directives or other incentives for the EU and member states’ governments and public administrations when it comes to purchasing AI and/or Language Technology products.
- Provide legislative incentives and/or regulations to make sure that all languages are covered (at least) in everyday services to the public. This includes TV and radio broadcasts that will not only be available across the border to a general audience, but will provide translation of any local or EU-wide programmes into all EU languages (as subtitles or dubbing) – perhaps by widening the scope of the Audiovisual Media Services Directive, originally targeted at the hearing-impaired. Similarly, to support smooth cross-border cooperation in all business as well as public affairs, all laws, regulations,

rules, etc., governing taxes, entrepreneurship, reporting rules, consumer rights, certifications, standardisation, and personal affairs, such as marriages, school systems, insurance affairs, home utilities, personal taxes, school rules and many more should be made multilingual (translated at least to all official EU languages), mandatory for everyone.

An ELE programme should be based on a long(er)-term vision, since the technology of 2030 will be eventually implemented in applications and services over the following decade, i. e. 2030-2040. Thus, while providing funding for closing the current gaps in data, efficiency of tools and services, and in helping businesses to thrive in their application areas, it should focus on future challenges. These include what is called today ‘General AI’, which – in the LT field – corresponds to ‘Natural Language Understanding’ (NLU). In this forward-looking perspective, NLU covers both written and spoken language, possibly in connection with vision, haptics and other senses inherent to humans, and in terms of behaviour, it can fully simulate/emulate humans in all possible situations.

This vision is partially complicated by the fact that Natural Language Understanding has no clear definition, apart from the Turing test. However, certain ‘component technologies’ that would provide building blocks for NLU systems may already be formulated as follows: one of the primary problems is to find the relation between the real world and language (in the general sense). Even philosophers do not agree whether our understanding of the world is only expressed by language or in fact formed by language; for practical purposes, it does not matter so much, but the relation is yet to be defined. For example, our current world knowledge is sparsely captured in various databases, such as DBpedia, Wikidata, domain ontologies (such as MESH, ICD, and others for the medical domain, or systematic classifications in biology or chemistry etc.), but – except perhaps for Wikipedia – there is no general world knowledge database or ontology which captures all that we know about our world. This is not a new problem – but it was confirmed by several partially successful attempts, such as CyC,¹⁵ that it is a very difficult task. Such a knowledge representation, whether built manually, from existing sources, or in some self-learning way, is crucial to relate and link documents, speeches and other language performances to the real world, in order to allow inferences (in a similar way that people do inferences), to allow for common sense conversations to solve common tasks between people and machines, to have a chat, to control various complex devices, respond to emergencies etc.

The ELE programme should then tackle, in its more future-oriented part, these issues:

- Design a general enough representation of the world around us, in cooperation with other AI programmes
- Build and/or assemble/convert existing world-knowledge databases to some instances of the common representation
- Design methods of inferencing, question answering, and constant updating of such knowledge base
- Propose a method for adapting or selecting the right subset for domain specific applications
- Develop data collection and/or efficient annotation of language data by such representation to allow for supervised learning or to test unsupervised machine learning approaches, annotation transfer between languages, domains, efficient reinforcement learning, etc., and to allow for more advanced cross-language and new learning techniques, including continuous learning; develop methods and algorithms to work with such representation and/or databases

¹⁵ <https://cyc.com>

- Develop efficient adaptation to applications, both in terms of language, domain, efficiency, power consumption, size and ease of maintenance, and quality assurance.

4.2.3. Legal and Administrative Support

At the level of policies or instruments, much more synchronisation of activities between national and international levels is necessary. For example, while all CLARIN member countries actively support international activities, including a membership fee, there is a significant gap between countries where national funding is concerned. Different speed in the implementation (and acceptance) of national roadmaps leads to a widening gap between languages (e. g. with respect to the Baltic countries, for many years Estonia had much better and more targeted support for national LT and CLARIN activities than Lithuania and Latvia (Skadina, 2018)).

Another important aspect is IPR regulation that needs to be more flexible, allowing wider use of IPR protected data for the development of language technologies and resources in a way that does not harm the interests of the authors.

An instrument for efficient and homogeneous implementation of DLE policy is equal support at international level through equal involvement of national research communities. Finally, DLE could be reached through international support for collaborative activities for research infrastructures aiming at language equality for some specific LT domain or area (e. g. similar to the CLARIN Resource Families Project funding, but on a larger scale).

With respect to translation technology, the key expectation is language equality, regardless of language pair, domain, complexity of language and availability of (training) data. The expectations include translation technology that ‘understands’ language, context and can use/is aware of common/grounded knowledge. Since translation technologies still face many challenges,¹⁶ an important aspect is sharing of data, knowledge and technologies through research infrastructures and LT platforms that support research and development activities, including collaboration, knowledge sharing, and open access to data and technologies.

The role of RIs will become even more essential in maximising the exploitation potential of publicly funded research results. This is one of the reasons why sufficient operational capacity of RIs needs to be ensured, so that RIs better can address the needs of their future professional users. Such needs will focus more than ever on comprehensive LT services and deep-learning language models.

4.2.4. Human Resources and Initiatives for Education

While by all accounts there are many capable and well-educated researchers available in Europe and new generations are constantly graduating from European colleges, it is necessary to provide incentives for them to stay in Europe, both for doctoral (PhD) studies and their post-academic career stages, both in education, research and in European industry. It is imperative that Europe can continue to educate the future generation of researchers and practitioners.

5. Conclusions

This report summarises the opinions that were collected within the CLARIN network on the matters of current status of LT in Europe and potential steps that are to be taken.

¹⁶ Multilinguality and Machine Translation are elaborated in the ELE deliverable *D2.13 Technology Deep Dive – Machine Translation*, available at <https://european-language-equality.eu/deliverables/>.

Looking into the future, the survey respondents and interviewees agree on a number of challenges that are to be dealt with in order to reach DLE by 2030, and to continue the steady development of NLU beyond this date.

First of all, the scientific development of LT should not only follow the needs of diverse groups of European citizens no matter how big or small their representative population is, but it should also be aligned to standardisation initiatives to harmonise and adopt the standards across different fields of applications. RIs working on building the EOSC are already laying the foundation to support smooth sharing of data, tools and services. While this work ensures the implementation of FAIR principles in all aspects of scientific work, this is where targeted EU funding can make a key difference to ensure an alignment on the national level of LT developments. At the same time, there should be a recognition of this kind of contributions as being part of evaluation and validation criteria. Further on, current cluster initiatives within domains (such as SSH), and further under the umbrella of EOSC, improve visibility and findability of resources and services across domains. Moreover, access to the data and tools via distributed RIs allows both to optimise the storage space and processing power, as well as to compare the LTs in regards to their computational footprint which is of crucial importance in order to deal with the issue of LT footprint.

Second, there is an agreement that more legal and administrative support for the field is a prerequisite for DLE. On the pan-European landscape there is no consistent way in which the European regulations are applied and funding is allocated, and this creates obstacles for researchers when accessing and further sharing data, and when preparing long-term scientific strategies. On the one hand, researchers need to have clear guidelines as to the application of the GDPR in their domain, not only on their national level, but also in the context of the international collaborations within Europe and beyond. A targeted campaign will help to drive further work on harmonising the application of European applications and, in turn, support international collaborations. On the other hand, the investment in LT should include appreciation and reward for the publications in local languages, for seeking solutions to the problems of smaller linguistic communities, for ensuring the reproducibility and trustworthiness of the research workflows and outcomes.

Third, human resources and attention to human experts are extremely important. High levels of educational standards that have already been achieved should be further strengthened by encouraging researchers, especially early career researchers, to get involved with real use cases and the plethora of available tools. A supportive scientific environment and appropriate funding are key to both ensuring the continuation of projects, as well as building and sharing of expertise knowledge through collaboration. While encouraging and supporting the (early career) researchers to address and solve the cutting-edge problems, this must go hand-in-hand with the allocation of resources. This allows fundamental research to continue and fosters a certain degree of freedom that is necessary in order to explore the challenges that may not be immediately visible on the surface, but which nonetheless have significant application value.

In the context of CLARIN's contribution to the work towards DLE, it can be stated that CLARIN's strategy¹⁷ is already aligned in various ways to the identified challenges. However, it is vital that the ongoing activities within the CLARIN consortium must be supported by large-scale funding to LT development at the European level outlined by the ELE programme.

References


Daan Broeder, David Nathan, Sven Strömqvist, and Remco Van Veenendaal. Building a federation of language resource repositories: the DAM-LR Project and its continuation within CLARIN. In *Sixth*

¹⁷ <https://www.clarin.eu/content/vision-and-strategy>

- International Conference on Language Resources and Evaluation (LREC 2008)*. ELRA, 2008.
- Franciska de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259 – 3264. ELRA, 2018. URL <https://dSPACE.library.uu.nl/handle/1874/364776>.
- Franciska de Jong, Bente Maegaard, Darja Fišer, Dieter van Uytvanck, and Andreas Witt. Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3406–3413. European Language Resources Association, May 2020. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.417>.
- Erhard Hinrichs and Steven Krauwer. The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1525–1531, 2014.
- Ondřej Košarko, Dušan Variš, and Martin Popel. LINDAT translation service, 2019. URL <http://hdl.handle.net/11234/1-2922>. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Mārcis Pinnis, Toms Bergmanis, Kristīne Metuzāle, Valters Šics, Artūrs Vasiļevskis, and Andrejs Vasiļjevs. A tale of eight countries or the EU council presidency translator in retrospect. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 525–546, Virtual, October 2020. Association for Machine Translation in the Americas. URL <https://www.aclweb.org/anthology/2020.amta-user.25>.
- Inguna Skadina. Some highlights of human language technology in baltic countries. In Audrone Lupeikiene, Olegas Vasilecas, and Gintautas Dzemyda, editors, *Databases and Information Systems X - Selected Papers from the Thirteenth International Baltic Conference, DB&IS 2018, Trakai, Lithuania, July 1-4, 2018*, volume 315 of *Frontiers in Artificial Intelligence and Applications*, pages 18–30. IOS Press, 2018. doi: 10.3233/978-1-61499-941-6-18. URL <https://doi.org/10.3233/978-1-61499-941-6-18>.
- Vésteinn Snæbjarnarson, Svanhvít Lilja Ingólfssdóttir, and Haukur Barri Símonarson. GreynirTranslate - mBART25 NMT (with layer drop) models for translations between icelandic and english, 2021. URL <http://hdl.handle.net/20.500.12537/128>. CLARIN-IS.
- Andrejs Vasiļjevs, Inguna Skadiņa, Indra Sāmīte, Kaspars Kauliņš, Ēriks Ajausks, Jūlija Meļņika, and Aivars Bērziņš. Competitiveness analysis of the European machine translation market. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 1–7, Dublin, Ireland, August 2019. European Association for Machine Translation. URL <https://aclanthology.org/W19-6701>.
- Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(160018), 2016. doi: 10.1038/sdata.2016.18. URL <http://dx.doi.org/10.1038/sdata.2016.18>.


A. The LT Researchers and Developers Full Survey

Figures 5 to 13 show the complete LT research and developers survey.



European Language Equality: Consultation with LT researchers and developers

Fields marked with * are mandatory.



EUROPEAN LANGUAGE EQUALITY

About this questionnaire

This questionnaire is delivered by the [European Language Equality \(ELE\)](#) project, a pilot action that addresses an appeal by the European Parliament resolution "[Language equality in the digital age](#)". The primary goal of ELE is to prepare a Strategic Research and Innovation Agenda and Roadmap, in order to tackle the striking imbalance between European languages in terms of the support they receive through **language technologies**.

To this end, ELE is reaching out to the European stakeholders involved in Digital Language Equality through a series of consultation rounds. This questionnaire is specifically addressed to **researchers and industry practitioners in the field of Language Technology (LT), Natural Language Processing (NLP), Speech Technologies and Language-centric AI**.

The questionnaire takes approximately 20 minutes to fill in. You are requested to evaluate the current situation with respect to the level of LT support for European languages, to indicate challenges and to share your needs and expectations for the future.

Your contributions will be carefully taken into account when preparing the ELE strategic agenda and roadmap.

This is a joint pan-European effort that will impact the field of LT in Europe for the next 10-15 years, including the funding situation. Join us and be a part of it!

Personal data protection

1

Figure 5: Full survey as published (page 1/9)

Personal data, i.e. name and email address, will be used **for contact purposes only** during the ELE project, i.e. to invite respondents to follow-up interviews or to the ELE conference or other project events. No personal data of the respondents will be made available to any third-party, beyond the ELE consortium. The names and emails of the respondents will not be reported in any project public document. The respondents' views and opinions, as expressed through this questionnaire, may be reported **anonymously** in the project's deliverables or in other public documents, e.g. scientific publications, dissemination material etc., without any reference to the individual's personally identifiable information.

Please read the [ELE Privacy policy](#) to get informed about the processing of your personal data when filling in this questionnaire.

1 Introduce yourself and your organisation

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

- University or other academic research organisation
- Research center (independent from any other academic organisation)
- SME
- Large enterprise
- Other

If "Other", please specify.

* What is the name of the organisation you work for?

If applicable, please provide the name of the LT-specific group within the organisation first, e.g. NLP group/Department of Linguistics /School of Philology/University of Athens.

* Where is your organisation's headquarters based?

- Austria
- Belgium
- Bulgaria
- Croatia
- Cyprus
- Czechia
- Denmark
- Estonia
- Finland
- France
- Germany
- Greece
- Hungary
- Iceland
- Ireland
- Italy
- Latvia
- Lithuania
- Luxembourg
- Malta
- Netherlands
- Norway
- Poland
- Portugal
- Romania
- Slovak Republic
- Slovenia
- Spain
- Sweden
- Other

If "Other", please specify.

2

Figure 6: Full survey as published (page 2/9)

* Which LT areas do you mainly work in?

- Basic natural language processing services (PoS tagging, parsing, named entity recognition etc.)
- Search and information retrieval technologies
- Text analytics and mining, information extraction, text classification
- Translation technologies (Machine Translation, translation memories management, CAT tools)
- Speech technologies
- Conversational systems
- Language resources: data production, data aggregation
- Language resources: data distribution, data marketplace
- Research infrastructures (e.g. catalogue, repository)
- Other

If "Other", please specify.

* Are you/your organisation a member of one or more of the following associations/networks/projects?

- CLARIN
- META-NET
- ELG
- CLAIRE
- LT-Innovate
- AI4EU
- ELEXIS
- BDVA
- AI PPP
- HumanE AI Network
- Nexus Linguarum
- ELISE
- TAILOR
- AI4Media
- VISION
- AI4Copernicus
- AIPlan4EU
- BonsAPPs
- DIH4AI
- I-ENERGY
- StairwAI
- Other
- None of the above

If "Other", please specify.

How many organisations participate in your national CLARIN consortium?

How many LT researchers/experts/students are employed and/or actively contribute to the national CLARIN consortium?

Please do not report the number of students using the resources in education only. Only the number of active contributors is relevant here.

In which sectors are your technologies, products or services used?

- Agriculture and fisheries
- Insurance industry

3

Figure 7: Full survey as published (page 3/9)

| | |
|---|---|
| <input type="checkbox"/> Digital Humanities, arts, culture and other services | <input type="checkbox"/> Justice and legal |
| <input type="checkbox"/> Broadcasting | <input type="checkbox"/> Media |
| <input type="checkbox"/> Business services | <input type="checkbox"/> Public administration |
| <input type="checkbox"/> Construction | <input type="checkbox"/> Publishing |
| <input type="checkbox"/> eCommerce | <input type="checkbox"/> Security (threat detection in general) |
| <input type="checkbox"/> Education | <input type="checkbox"/> Social Sciences |
| <input type="checkbox"/> Energy/green economy/environment | <input type="checkbox"/> Tourism, accommodation and food services |
| <input type="checkbox"/> Finance/banking | <input type="checkbox"/> Trade and repair |
| <input type="checkbox"/> Health | <input type="checkbox"/> Transportation, logistics and storage |
| <input type="checkbox"/> Industry and manufacturing | <input type="checkbox"/> Other |
| <input type="checkbox"/> Information and Communication Technologies | |

If "Other", please specify.

2 Language coverage

*What languages does your organisation conduct research in and/ or for what languages do you offer services, software, resources, models etc.?

| | | |
|---|--|-------------------------------------|
| <input type="checkbox"/> Basque | <input type="checkbox"/> Galician | <input type="checkbox"/> Norwegian |
| <input type="checkbox"/> Bulgarian | <input type="checkbox"/> German | <input type="checkbox"/> Polish |
| <input type="checkbox"/> Catalan; Valencian | <input type="checkbox"/> Greek | <input type="checkbox"/> Portuguese |
| <input type="checkbox"/> Croatian | <input type="checkbox"/> Hungarian | <input type="checkbox"/> Romanian |
| <input type="checkbox"/> Czech | <input type="checkbox"/> Icelandic | <input type="checkbox"/> Serbian |
| <input type="checkbox"/> Danish | <input type="checkbox"/> Irish | <input type="checkbox"/> Slovak |
| <input type="checkbox"/> Dutch | <input type="checkbox"/> Italian | <input type="checkbox"/> Slovenian |
| <input type="checkbox"/> English | <input type="checkbox"/> Latvian | <input type="checkbox"/> Spanish |
| <input type="checkbox"/> Estonian | <input type="checkbox"/> Lithuanian | <input type="checkbox"/> Swedish |
| <input type="checkbox"/> Finnish | <input type="checkbox"/> Luxembourgish | <input type="checkbox"/> Welsh |
| <input type="checkbox"/> French | <input type="checkbox"/> Maltese | <input type="checkbox"/> Other |

If "Other", please specify.

Please separate multiple languages with a comma (,).

Are there any languages that your organisation does not yet support, but you plan to support in the next three years?

| | | |
|---|------------------------------------|-------------------------------------|
| <input type="checkbox"/> Basque | <input type="checkbox"/> Galician | <input type="checkbox"/> Norwegian |
| <input type="checkbox"/> Bulgarian | <input type="checkbox"/> German | <input type="checkbox"/> Polish |
| <input type="checkbox"/> Catalan; Valencian | <input type="checkbox"/> Greek | <input type="checkbox"/> Portuguese |
| <input type="checkbox"/> Croatian | <input type="checkbox"/> Hungarian | <input type="checkbox"/> Romanian |
| <input type="checkbox"/> Czech | <input type="checkbox"/> Icelandic | <input type="checkbox"/> Serbian |
| <input type="checkbox"/> Danish | <input type="checkbox"/> Irish | <input type="checkbox"/> Slovak |
| <input type="checkbox"/> Dutch | <input type="checkbox"/> Italian | <input type="checkbox"/> Slovenian |

4

Figure 8: Full survey as published (page 4/9)

English
 Estonian
 Finnish
 French

Latvian
 Lithuanian
 Luxembourgish
 Maltese

Spanish
 Swedish
 Welsh
 Other

If "Other", please specify.
 Please separate multiple language with a comma (,).

* Considering your development plans with respect to language coverage, what are the **top three** drivers for your decision to support additional languages?
at most 3 choice(s)
 Please choose a maximum of 3.

- Market interest/demand by users or customers
- Research/scientific interest
- Available funding/investment
- Availability of human experts for other languages
- Availability of language resources
- Availability of technologies/tools
- Other

If "Other", please specify.

3 Evaluation of current situation

Please indicate if you agree with the following statements: **"One of the main challenges and obstacles the European LT community currently faces is..."**

| | Strongly agree | Agree | Disagree | Strongly disagree | <i>I Don't know / No answer</i> |
|---|-----------------------|-----------------------|-----------------------|-----------------------|---------------------------------|
| * ...basic research is still needed." | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| * ...inadequate recognition of the importance of multilinguality." | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| * ...lack of talent/brain drain." | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| * ...fragmentation of the European LT industry." | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| * ...lack of coordination and missing links between research, LT vendors, integrators and customers." | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| * ...insufficient public procurement." | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

5

Figure 9: Full survey as published (page 5/9)

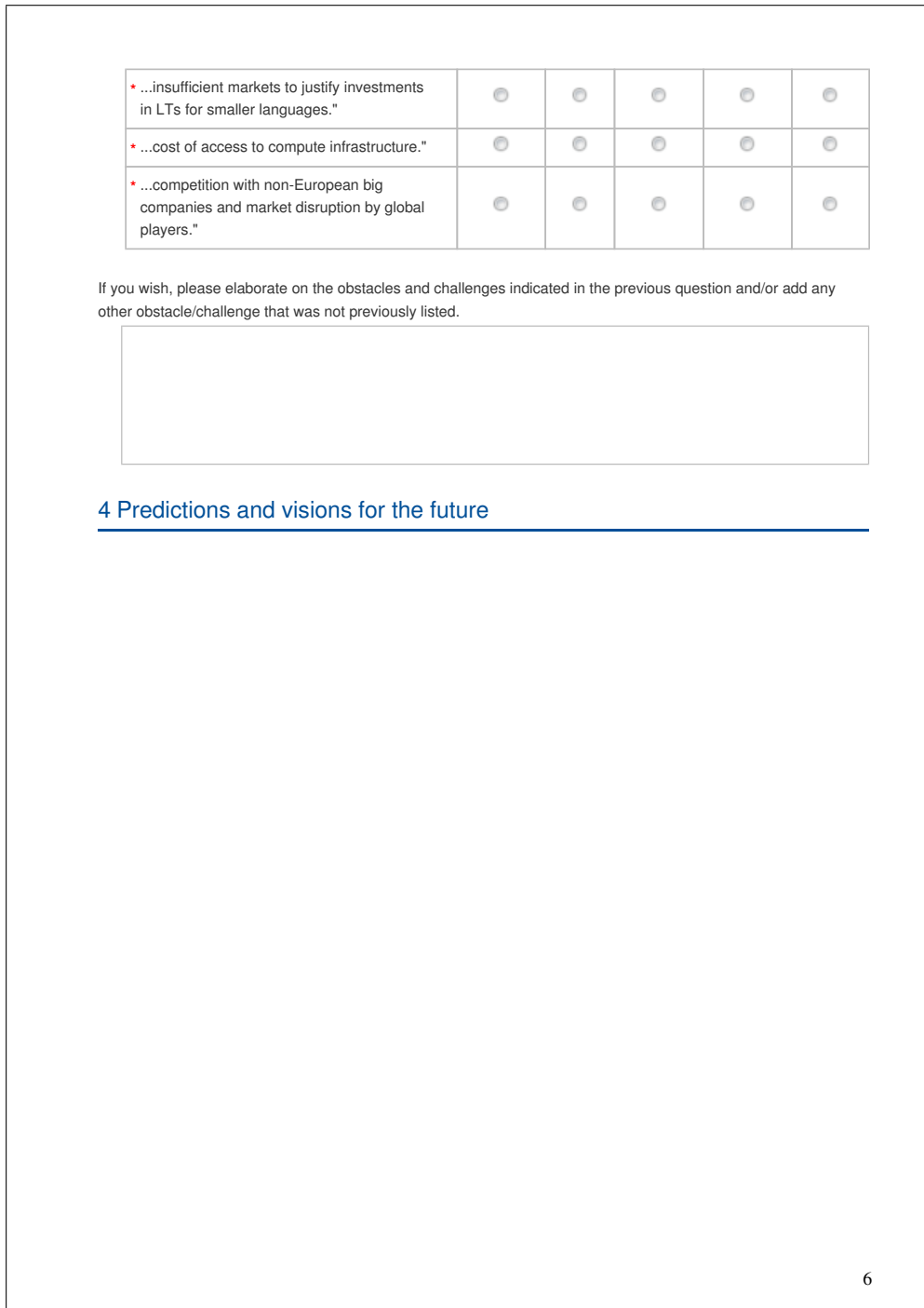


Figure 10: Full survey as published (page 6/9)

In your opinion, how effective can the following policies/instruments be in speeding up the development and deployment of LT in Europe equally for all languages?

| | Very effective | Effective | Moderately effective | Slightly effective | Not effective at all | <i>/ don't know / No answer</i> |
|--|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|---------------------------------|
| • Initiate large-scale, long-term funding programme for European LT development | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| • Initiate investment instruments and accelerator programs targeting LT start-ups | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| • Continuous investment in the Research Infrastructures that support LT. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| • Increase availability of qualified personnel on LT and incentives for talent retention | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| • Public procurement of innovative technology and pre-commercial public procurement | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

7

Figure 11: Full survey as published (page 7/9)

| | | | | | | |
|---|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| • Raise awareness of the benefits for companies, public bodies, and citizens of the availability of on-line services, contents and products in multiple languages | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| • Impose content accessibility regulations, e. g., multimedia subtitling, readability, dubbing, availability of content in multiple languages etc. | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| • Invest in the development of new (scientific /technological) methodologies for transfer /adaptation of resources /technologies to other domains and languages | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| • Reinforce training and education initiatives, including undergraduate and masters programs and vocational training in LT | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

8

Figure 12: Full survey as published (page 8/9)

Are there any other policies/instruments not listed in the previous question, which in your opinion can be effective be in speeding up the development and deployment of LT in Europe equally for all languages?

If there is a large-scale, long-term funding programme dedicated to European Language Technology research, development and innovation running for approx. ten years, what are, in your opinion, the **(up to) five key challenges** Europe needs to concentrate on with regard to basic and applied research?

If there is a large-scale, long-term funding programme dedicated to European Language Technology research, development and innovation running for approx. ten years, what are, in your opinion, the **(up to) five key challenges** Europe needs to concentrate on with regard to **innovation and the LT industry**?

Do you have any other additional suggestions or recommendations with regard to European Language Equality?

Can we contact you to arrange a possible follow-up discussion?

- Yes
 No

*What is your email address?

What is your name?

By clicking on 'Submit', I agree that my personal data (email address and/or name) can be used according to the Privacy Policy of the European Language Equality (ELE) project.

[ELE Privacy Policy.pdf](#)

Figure 13: Full survey as published (page 9/9)

B. Additional Tables and Graphs

| Type of organisation | Answers (Perc.) | |
|--|-----------------|-----|
| University or other academic research organisation | 77 | 86% |
| Research center (independent from any other academic organisation) | 9 | 10% |
| SME | 2 | 2% |
| Other | 2 | 2% |
| Total | 90 | |

Table 3: Breakdown of answers to “Which of the following best describes the type of organisation you work for?” (mandatory closed question)

| Country | Respondents (Perc.) | |
|--------------------|---------------------|-----|
| Czechia | 10 | 11% |
| Greece | 9 | 10% |
| Sweden | 8 | 9% |
| Denmark | 6 | 7% |
| Poland | 5 | 6% |
| Slovenia | 5 | 6% |
| Italy | 4 | 4% |
| Portugal | 4 | 4% |
| Spain | 4 | 4% |
| Belgium | 3 | 3% |
| Finland | 3 | 3% |
| Germany | 3 | 3% |
| Latvia | 3 | 3% |
| Lithuania | 3 | 3% |
| Austria | 2 | 2% |
| Bulgaria | 2 | 2% |
| Croatia | 2 | 2% |
| Estonia | 2 | 2% |
| Hungary | 2 | 2% |
| Netherlands | 2 | 2% |
| United Kingdom | 2 | 2% |
| Cyprus | 1 | 1% |
| France | 1 | 1% |
| Iceland | 1 | 1% |
| <i>Luxembourg</i> | 1 | 1% |
| Norway | 1 | 1% |
| <i>Switzerland</i> | 1 | 1% |
| Total | 90 | |

Table 4: Breakdown of answers to “Where is your organisation’s headquarter based?” (mandatory closed question, plus “if other” as optional open-ended question). The countries that are not CLARIN members are marked in Italics.

| Sector | Answers |
|--|----------------|
| Digital Humanities, arts, culture and other services | 78 |
| Education | 62 |
| Information and Communication Technologies | 55 |
| Social Sciences | 45 |
| Media | 30 |
| Health | 27 |
| Public administration | 22 |
| Justice and legal | 18 |
| Business services | 16 |
| Broadcasting | 16 |
| Finance/banking | 11 |
| Publishing | 11 |
| Industry and manufacturing | 7 |
| Tourism, accommodation and food services | 6 |
| eCommerce | 5 |
| Insurance industry | 3 |
| Energy/green economy/environment | 3 |
| Transportation, logistics and storage | 3 |
| Security (threat detection in general) | 3 |
| Agriculture and fisheries | 2 |
| Construction | 1 |
| Other (Lexicography) | 1 |
| Other (Linguistics) | 1 |
| Other (Music industry) | 1 |
| Other (Sworn translators and interpreters) | 1 |

Table 5: Breakdown of answers to “In which sectors are your technologies, products or services used?” (mandatory closed question, plus “if other” as optional open-ended question).

| Language | In work (Q14) | Planned in 3 years (Q16) |
|---------------|---------------|--------------------------|
| Basque | 6 | 1 |
| Bulgarian | 10 | 2 |
| Catalan | 6 | 1 |
| Croatian | 12 | 1 |
| Czech | 21 | 1 |
| Danish | 15 | - |
| Dutch | 16 | - |
| English | 65 | 4 |
| Estonian | 10 | 1 |
| Finnish | 7 | 1 |
| French | 22 | 4 |
| Galician | 6 | 1 |
| German | 25 | 4 |
| Greek | 19 | - |
| Hungarian | 12 | 1 |
| Icelandic | 8 | 1 |
| Irish | 3 | 1 |
| Italian | 21 | 1 |
| Latvian | 7 | 1 |
| Lithuanian | 8 | 3 |
| Luxembourgish | 2 | 1 |
| Maltese | 3 | 1 |
| Norwegian | 8 | - |
| Polish | 15 | 1 |
| Portuguese | 14 | 1 |
| Romanian | 8 | - |
| Serbian | 9 | 1 |
| Slovak | 9 | 2 |
| Slovenian | 14 | 1 |
| Spanish | 20 | 2 |
| Swedish | 19 | 1 |
| Valencian | 6 | 1 |
| Welsh | 3 | 1 |

Table 6: Breakdown of answers to questions Q14 and Q16 “What languages does your organisation conduct research in and/ or for what languages do you offer services, software, resources, models etc.?” and “Are there any languages that your organisation does not yet support, but you plan to support in the next three years?” respectively.

| Statement | Strongly agree | Agree | Disagree | Strongly disagree | I don't know / No answer |
|--|----------------|-----------|----------|-------------------|--------------------------|
| basic research is still needed | 44 | 38 | 4 | 1 | 3 |
| inadequate recognition of the importance of multilinguality | 25 | 41 | 17 | 2 | 5 |
| lack of talent/brain drain | 10 | 27 | 31 | 11 | 11 |
| fragmentation of the European LT industry | 12 | 42 | 10 | 2 | 24 |
| lack of coordination and missing links between research, LT vendors, integrators and customers | 11 | 45 | 14 | 3 | 17 |
| insufficient public procurement | 14 | 34 | 14 | 4 | 24 |
| insufficient markets to justify investments in LTs for smaller languages | 25 | 40 | 26 | 2 | 7 |
| cost of access to compute infrastructure | 8 | 37 | 27 | 4 | 14 |
| competition with non-European big companies and market disruption by global players | 24 | 38 | 13 | 3 | 12 |

Table 7: Answers to the question (Q20-Q28): “Please indicate if you agree with the following statements: “One of the main challenges and obstacles the European LT community currently faces is...” (mandatory closed question, answers provided on a four-point scale, plus “I don’t know/No answer”). The statements and numbers in bold represent the answers where the audience predominantly (more than 70 percent) agrees with the statement.

| Statement | Very effective | Effective | Moderately effective | Slightly effective | Not effective at all | I don't know / No answer |
|---|-----------------------|------------------|-----------------------------|---------------------------|-----------------------------|---------------------------------|
| Continuous investment in the Research Infrastructures that support LT | 53 | 30 | 4 | 2 | 0 | 1 |
| Initiate large-scale, long-term funding programme for European LT development | 47 | 26 | 8 | 4 | 1 | 4 |
| Reinforce training and education initiatives, including undergraduate and masters programs and vocational training in LT | 34 | 28 | 19 | 4 | 1 | 4 |
| Public procurement of innovative technology and pre-commercial public procurement | 34 | 14 | 19 | 8 | 1 | 14 |
| Increase availability of qualified personnel on LT and incentives for talent retention | 28 | 39 | 15 | 1 | 1 | 6 |
| Invest in the development of new (scientific/technological) methodologies for transfer/adaptation of resources/technologies to other domains and languages | 28 | 35 | 17 | 5 | 2 | 3 |
| Raise awareness of the benefits for companies, public bodies, and citizens of the availability of on-line services, contents and products in multiple languages | 25 | 26 | 25 | 5 | 3 | 6 |
| Initiate investment instruments and accelerator programs targeting LT start-ups | 21 | 31 | 23 | 2 | 5 | 8 |
| Impose content accessibility regulations, e.g., multimedia subtitling, readability, dubbing, availability of content in multiple languages etc. | 19 | 30 | 21 | 9 | 3 | 8 |

Table 8: Answers to the question (Q30-Q38): “In your opinion, how effective can the following policies/instruments be in speeding up the development and deployment of LT in Europe equally for all languages?” (mandatory closed question, answers provided on a five-point scale, plus “I don’t know/No answer”).