EUROPEAN
LANGUAGE
EQUALITY

D3.3

Report on the final round of feedback collection

Authors	Itziar Aldabe (UPV/EHU), Aritz Farwell (UPV/EHU), German Rigau (UPV/EHU)
Dissemination level	Public
Date	31-05-2022

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	D3.3
Deliverable title	Report on the final round of feedback collection
Type	Report
Number of pages	31
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 31-05-2022 – Actual: 31-05-2022
Work package	WP3: Development of the Strategic Agenda and Roadmap
Task	Task 3.3 Final round of feedback collection
Authors	Itziar Aldabe (UPV/EHU), Aritz Farwell (UPV/EHU), German Rigau (UPV/EHU)
Reviewers	Stelios Piperidis (ILSP), Georg Rehm (DFKI)
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2022 ELE 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	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, 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	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (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

List of Acronyms

AI	Artificial Intelligence
ASR	Automated Speech Recognition
CLAIRE	Confederation of Laboratories for AI Research in Europe
CLARIN	Common Language Resources and Technology Infrastructure
CLTL	Cross-Lingual Transfer Learning
DLE	Digital Language Equality
EC	European Commission
ECSPM	European Civil Society Platform for Multilingualism
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELG	European Language Grid (EU project, 2019-2022)
EU	European Union
HPC	High-Performance Computing
LIBER	Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries
LR	Language Resources/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
MRLUs	Minority/Regional/Lesser-Used Languages
MT	Machine Translation
NEM	European Institute for Research and Strategic Studies in Telecommunications GmbH
NGO	Non-Governmental Organization
NLP	Natural Language Processing
NLU	Natural Language Understanding
R&D	Research and Development
RML	Regional and Minority Language
SRIA	Strategic Research and Innovation Agenda
WP	Work Package

Abstract

The primary objective of the ELE project is to prepare the European Language Equality Programme in the form of a Strategic Research and Innovation Agenda (SRIA), as well as a roadmap for achieving full Digital Language Equality (DLE) in Europe by 2030. This deliverable reports on the insights gained from Task 3.3 ("Final round of feedback collection"). It summarizes the methodology and meetings conducted to collect feedback from consortium partners and relevant stakeholders to improve the draft SRIA. This was completed during two meetings on May 6th and May 13th, 2022, in which the main findings from WP1, WP2 and WP3 were presented and recommendations were received from the consortium. These assisted in consolidating ELE's strategic plan to advance LT and language-centric AI research, technology, infrastructure and policy. The proposals gathered here will aid in designing and establishing the shared European programme for Language Technology and Digital Language Equality.

1 Introduction

This deliverable reports on the insights gained from Task 3.3 ("Final round of feedback collection"). We present the methodology and meetings conducted to collect the final round of feedback regarding the Strategic Research and Innovation Agenda (SRIA). The aim of this process was to collect feedback to improve the draft SRIA as well as to include the main conclusions that all consortium partners reached. ELE brings together a large number of partners representing many LT areas and natural languages relevant for the development of the SRIA. Over the course of two meetings, brief reports on the main findings of Work Packages 1-3 were presented and feedback was received in the form of recommendations for the individual languages as well as the final SRIA. The overview of WP1 outlined the process through which the definition of Digital Language Equality was reached, the ELE report on the state of the art in LT and language-centric AI, the 34 language reports, and the creation of the ELE/ELG Dashboard. The review of WP2 summarised the results derived from the surveys conducted on LT developers and users, as well as the deep dives into key technology areas. The synopsis of WP3 discussed the development of the strategic agenda and road map, including the report on existing strategic documents and projects.

In the following, we briefly describe the methodology followed to present the project's primary conclusions, together with the results produced from the feedback collection process.

2 Methodology

In order to obtain feedback from all ELE partners regarding the principal outcomes of the project, two online meetings were organized, one on May 6th and another on May 13th, 2022. The purpose of the first meeting was to receive feedback from all consortium members and relevant stakeholders as well as to provide an overview of the results of the SRIA and roadmap to date. A summary was given of the methodology and main findings from WP2 and D2.18, "Report on the state of Language Technology in 2030" (Way et al., 2022). In addition, the final conclusions drawn from the report on the "State of the Art in Language Technology and Language-centric AI" (D1.2, Agerri et al., 2021) and the report on existing strategic documents and projects in LT/AI (D3.1, Aldabe et al., 2021) were presented. Consortium partners were invited to provide feedback to the presentations as well as their perspectives concerning the recommendations to be included in the SRIA.

For the second meeting, all consortium partners, and in particular those involved in producing the language reports[1] and representatives of the LT community who participated on the surveys[2], contributed feedback to the aforementioned presentations and stated their primary conclusions, drawing from their work on the language reports. Their responses contained recommendations for the individual languages as well as the SRIA, including ideas they considered especially relevant or that need stronger emphasis. The two meetings resulted in a series of recommendations that are summarised in the following sections.

3 First Meeting

Sixty-three consortium members participated in the first meeting. As previously noted, the purpose of the meeting was to highlight some of ELE's principal findings as well as to introduce the first version of the draft SRIA, the evidence-based strategic research agenda and road map to ensure that digital language equality becomes a reality in the EU by 2030.

First, WP3 was summarised and Tasks 3.1-3.3 were discussed. For Task 3.1, over 200 relevant documents and research papers were examined (Aldabe et al., 2021). Task 3.2 consolidates all input received from the 53 deliverables. Task 3.3, the final round of feedback collection, is incorporated into D3.3 (this deliverable). Second, a summary of the methodology and main findings from WP2 and D2.18, "Report on the state of Language Technology in 2030" (Way et al., 2022), was presented. WP2 analysed two sets of material. One came from the surveys: the LT developers survey, the LT users and consumers survey, and the EU citizen survey. The other was derived from the deep dives, which prompted industrial representatives in key technology areas to encapsulate present capabilities and to speculate about the future.

The LT developers survey received over 320 responses, representing 223 organizations across 32 countries. The surveys were distributed through META-NET, CLAIRE, CLARIN, LT Innovate and ELG. Close to three quarters of respondents were from the academic world, while just over twenty percent came from industry. The general results, broadly speaking, demonstrate that language resources were the main concern, especially with respect to smaller languages. The need for greater basic research into Natural Language Understanding (NLU) was another common theme. Many respondents also stressed that more collaboration between the EU and national centres is required. For some, this touched on the need for talent retention. Europe currently possesses adequate talent and education in AI, but experts continue to be lost to other countries.

The LT users and consumers survey received almost 250 responses and was conducted through organizations on the user side, including ECSPM, EFNIL, ELEN, LIBER, NEM and Wikimedia DE. The majority of respondents were from academia, but NGOs, government institutions, and companies also provided input. There were four general results, which overlapped somewhat with the developers survey. The first is that users wished to see more tools and resources developed for their languages. Some of the specific tools that were mentioned fell under the categories of MT, proofing tools, search engines and language learning. Respondents also provided feedback on perceived gaps. Among the most common was a lack

[1] Sarasola et al. (2022); Koeva and Stefanova (2022); Melero et al. (2022a); Tadić (2022); Hlavacova (2022); Pedersen et al. (2022); Steurs et al. (2022); Maynard et al. (2022); Muischnek (2022); Lindén and Dyster (2022); Adda et al. (2022); Sánchez and Mateo (2022); Hegele et al. (2022a); Gavriilidou et al. (2022); Jelencsik-Mátyus et al. (2022); Rögnvaldsson (2022); Lynn (2022); Magnini et al. (2022); Skadiņa et al. (2022); Gaidienė and Tamulionienė (2022); Anastasiou (2022); Rosner and Borg (2022); Eide et al. (2022); Ogrodniczuk et al. (2022); Branco et al. (2022); Păiș and Tufiș (2022); Garabík (2022); Krek (2022); Melero et al. (2022b); Borin et al. (2022); Prys et al. (2022); Krstev and Stanković (2022); Ćušić (2022); Moshagen et al. (2022)

[2] Thönnissen (2022); Eskevich and de Jong (2022); Rufener and Wacker (2022); Hajič et al. (2022); Hegele et al. (2022b); Gísladóttir (2022); Kirchmeier (2022); Hicks (2022); Blake (2022); Hrasnica (2022); Heuschkel (2022)

of tool variety for a given language. Basic ASR services, for example, are often not available for many languages. Taken together, this results in uneven language coverage.

The EU citizen survey was not originally planned in the ELE project. Nonetheless, in order to cast a wider net across Europe, the survey was carried out with the help of a commercial market research provider (Lucid). Over 21,000 responses from 31 countries were received. Preliminary findings from the survey were in line with those mentioned above. Once again, respondents not only asked for more tools and resources for their respective languages, including MT, search engines, and proofing tools, but also expressed a desire for personal assistants due to a lack of availability in many languages. In addition, respondents indicated that English, German and French were the languages they employed most online.

The language deep dives (Bērziņš et al., 2022; Backfried et al., 2022; Gomez-Perez et al., 2022; Kaltenboeck et al., 2022) took into account recommendations and predictions from almost all partners on the industrial side. They were divided into four technological areas: 1) machine translation, 2) text analytics and natural language understanding, 3) speech technologies and 4) data, language resources and knowledge graphs (across all LTs).

The vision for machine translation in 2030 (Bērziņš et al., 2022) includes seamless and ubiquitous translation availability for both text and speech. To get there, several hurdles will need to be surmounted, mostly through additional research. Among these are: better awareness of context and the ability to consider metadata, output that is faithful to the intended purpose of communication, and the capacity to explain text rather than simply translating, reflecting the possible cultural differences between the source and target languages. Similarly, speech translations will need to be able to show emotions when necessary or appropriate.

The vision for speech technology in 2030 (Backfried et al., 2022) foresees a world where speech input is enabled for most applications, environments and use scenarios. Several milestones must be reached through research in this case as well. Close to perfect performance is needed, for instance, because current speech technologies still do not do well in noisy or otherwise difficult environments. To help with this issue, data from these kinds of environments is a must. Additionally, models that combine speech signals with text and other modalities are needed in order to provide additional information. There are also several submodules that help with the integration of speech into applications (advanced speaker identification, speech diarization, multi-speaker ASR). Finally, more research into sign languages needs to be encouraged.

The vision for text analytics and NLU in 2030 (Gomez-Perez et al., 2022) forecasts the inclusion of deep learning and symbolic methods into various applications. Research will need to improve knowledge extraction from text, transcripts and multimodal input. Structured databases and knowledge graphs must be better linked to unstructured texts. In addition, multi-language models are required to work on all languages simultaneously and transparently. Lastly, ways to integrate NLU into LTs to improve accuracy and natural communication must be investigated.

The vision for data, language resources and knowledge graphs in 2030 (Kaltenboeck et al., 2022) confirms that data will continue to play a crucial part in developing LT. Two paths forward are needed in this regard. On the one hand, more data must be collected for novel applications. On the other, research would do well to discover how to utilize less data for applications while maintaining quality. Relatedly, power consumption in machine learning must be reduced. This is not only important for the environment, but also as a means to avoid the need for HPCs.

3.1 Recommendations

A handful of recommendations were presented along with the main findings from WP2. We list these below in Section 3.1.1, along with several others in Section 3.1.2 that were offered as part of the discussion that followed.

3.1.1 Recommendations based on WP2

Taken together, there are several key technological recommendations. At least five recommendations should be put in place in order to implement the visions outlined for the LT areas:

- Basic research must be supported for deep learning, neurosymbolic and other approaches to NLU.
- The speech and NLP community should join in efforts to create integrated models for all applications.
- Speech applications need to be adjusted to overcome the difficulties they demonstrate when engaging with real-world environments and speakers' idiosyncrasies.
- Large pretrained multi-language models for LTs must be fashioned because these usually do not require comparable power.
- In the case of MT, support is needed for integration with speech for real-time, multi-agent and multi-language "instant" spoken MT among all EU languages.

Four key recommendations were made with respect to data.

- Access to HPCs should be increased for all, but especially for academic institutions.
- The availability of data must also be increased.
- The energy footprint needs to be minimized.
- Legal conditions must be adjusted accordingly in order to reuse data in research.

3.1.2 Recommendations based on discussion

- Caution should be taken not to over promise what LT and AI will be able to deliver in the coming years. There is a danger, if unrealistic expectations are not met, that people interpret failure as an indication that the technology cannot succeed or that it has certain inherent limitations. Expressing realistic goals should be taken into account when engaging with the media.
- There are three keywords, or ideas, that should be emphasized with respect to MT. One is culture. MT must be able to consider context and cultural issues. The second keyword is emotions. Research into identifying or generating emotions should be encouraged. The third keyword is indigenous languages. Approaches to working on languages with fewer resources that do not rely on the large machine-learning paradigm should be developed.
- Research should not solely focus on language models and needs to be kept open to novel ideas to ensure that approaches which may differ from neural networks are not dismissed out of hand.

- Research may be aided by establishing a well-defined taxonomy of communication scenarios. This may help with increasing sensitivity to different kinds of communicative situations or scenarios. To tackle the variety of communication scenarios in which people find themselves, it is necessary to classify them so that they may be worked on independently.
- Applied linguistics may be able to assist when classifying reliable communication taxonomies. This is another example of why greater interdisciplinary collaboration and approaches are fundamental.
- Users need to be brought into the process to ensure that the tools and datasets that are built are useful to those who use them on a daily basis.
- Languages that are in dire need of digital support in terms of tools and data require support or there is a danger that they might be lost within the digital sphere. This problem would be aided in part by targeted calls to build resources or datasets for particular languages.
- There is a strong need for European coordination of LT because the situation in various countries in terms of language and technology funding is quite heterogeneous. There should be an umbrella approach at the European level, where the EU or the EC supports the overall topic of LT development, which is then complemented at the national level.
- Simultaneously, it is imperative that individual countries do their part to develop resources for their languages in a coordinated fashion with the EU. A multi-pronged approach to tool and resource building in which local or regional funding is sought at the same time as European funding would be ideal. The DLE metric Gaspari et al. (2022) can be utilized as a means to underscore the urgency to act and the respective priorities that must be fulfilled.
- Greater collaboration and communication within the LT community is important because more integration and more concerted action with respect to calls creates a louder and clearer message on what is needed.

4 Second Meeting

The second meeting began with an overview of WP1 with presentations regarding the definition of Digital Language Equality (Gaspari et al., 2022) and the creation of the ELE/ELG Dashboard (Giagkou et al., 2022). Additionally, it featured short presentations providing feedback and recommendations on the SRIA or on particular languages from various members of the consortium. Participants were asked to provide feedback in a short bulleted list (3-6 suggestions). This feedback contained, for example, emphasis on ideas that are considered especially relevant or received insufficient attention. Over 65 consortium members participated in this second meeting and 31 presentations were given. Figure 1 depicts a word cloud based on the feedback received. The size of each word indicates its importance based on the frequency. Thus, the most frequent terms cover aspects such us language, funding, research, data and tools.

4.1 Recommendations based on Feedback Collection

This section presents the recommendations for the SRIA based on the feedback collected from the consortium partners. For the complete list of received feedback consult Appendix A.

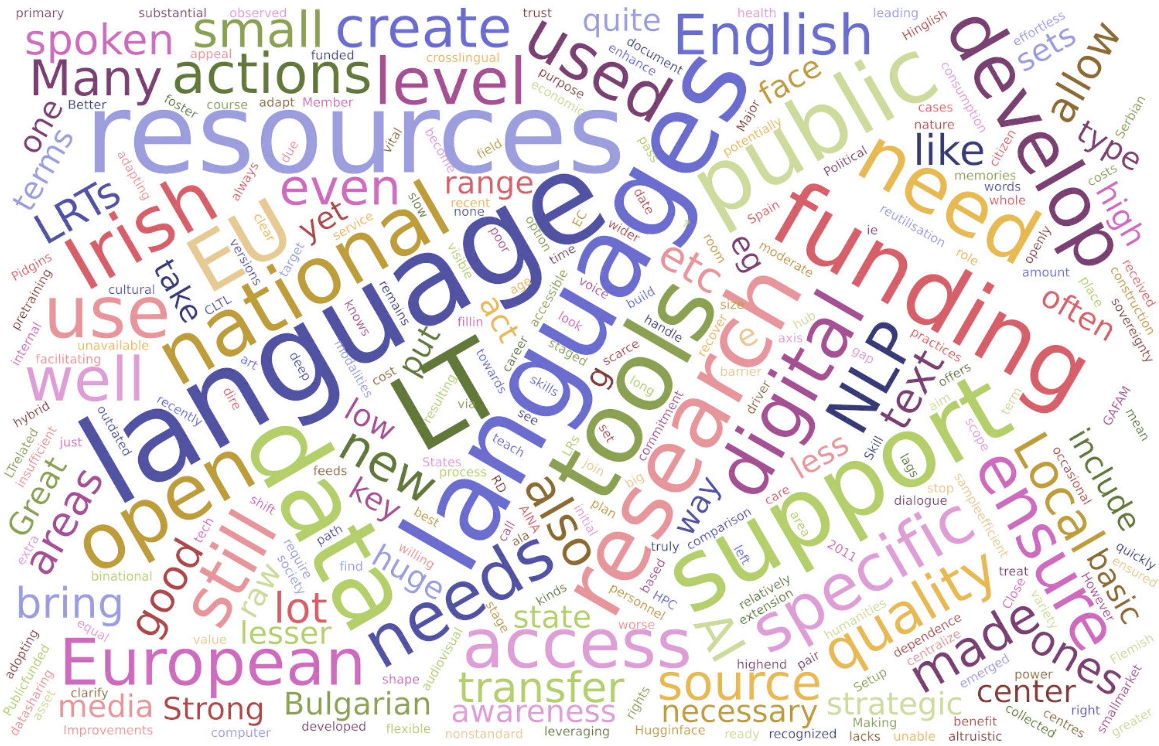

Figure 1: Word cloud corresponding to the feedback collected

The recommendations have been divided into SRIA and language-specific recommendations, as this was how feedback was obtained. In addition, they are grouped into specific categories: research, technology and data, infrastructure and policy. The SRIA recommendations are presented first, followed by their language counterparts.

4.1.1 SRIA Recommendations

Research Recommendations

Research recommendations for less-resourced languages include a call for novel techniques that would bring these to a level comparable to state-of-the-art results for resource-rich languages, as well as for their inclusion in large-scale multinational and multilingual R&D programmes of the type previously reserved for official EU languages. Research recommendations for LT more generally not only include placing greater emphasis on the crucial role LT plays within intelligent interaction, knowledge management, trust, and conflict resolution, but also note that more work should be invested in algorithms that include explainability, easy error correction, guaranteed performance, and knowledge ingestion and extensibility.

Technology and Data Recommendations

- Several recommendations touch upon language models. Large-scale language resources that can power language models with wide-ranging applications must be developed through coordinated action. The size, availability and quality of raw corpora that are capable of training language models should be augmented. Europe requires greater high-performance computing in order to boost AI and NLP, key for the ability to develop standard optimized language models. There is a need for general-purpose language-centric AI models that are trained on cross-language and cross-domain resources. These can benefit from adaptation to local language varieties and specialized domains with small or medium-sized datasets.

- The need to facilitate open-source solutions and language data sharing ranks among the top of all recommendations. This includes raising awareness about the importance of language data, increasing the availability of open-source material and promoting a culture of data sharing that involves stakeholders, the public sector, research and industry. Doing so would prove especially beneficial to languages with fewer speakers.
- An effort must be made to increase multimedia LT resources and further extend LT to a diversity of domains.
- Improvements to LT evaluation through more annotated benchmark corpora and multilingual benchmark datasets were suggested. Similarly, others believe standards for user-driven quality assessment should be developed.
- There is a demand for standardized workflows for annotated corpora generation and support.
- Low-resource languages would benefit from multilingual and multimedia data at the European level. They also require more translation tools.

Infrastructure Recommendations

LT-related infrastructures such as ELG must be maintained and extended. Ideally, these infrastructures would rely on close synchronization between national and international research objectives. Similarly, European and national coordinated actions are needed to ensure access to open high-performance computing infrastructures.

Policy Recommendations

- There is a demand for long-term funding for projects and institutions working with regional languages. More generally, the EC should recover its previous vision concerning LT, which included ambitious and targeted funding for research, development and innovation.
- The EU and its member states must demonstrate a political commitment to facilitate a path towards the reuse of language data. There is widespread agreement that policies are needed which guarantee open access to data. Regulations governing ready-to-use datasets based on public data from European, national, regional, and local public institutions would be helpful. As would legislation that clarifies the legal stance on reuse of data in Europe. This is especially true for data usage rights that facilitate NLP research and development.
- Put in place initiatives to create multi-domain databases and resources for low-resourced languages.
- Additional support must be given to low-resourced languages, especially in areas where LT is not market-driven. Such languages need individual funding programmes to develop basic language tools and resources that will help ensure their digital survival. The dire situation of Minority/Regional/Lesser-Used languages (MRLUs) cannot be improved without EU and national funding specifically dedicated to digital support. National and EU policy makers must be made aware of the need for this support.

- It is important to create opportunities for speakers of MRLUs to study areas related to LT and, in general, to offer more NLP courses to students as a way to strengthen NLP's viability as a career option. This might include, for instance, specific programmes to train and up-skill those working with endangered languages.
- Collaboration between academia and industry should be encouraged and it may be helpful to develop exchange programmes as a means to share knowledge.
- More must be done to convince the EU and member states that LT is a key geopolitical asset, especially in those places where multilingualism is a vital challenge.

4.1.2 Language related Recommendations

Research Recommendations

Research recommendations were varied. They include the following suggestions:

- Many resources that are currently in a "proof of concept" stage should be brought into usability.
- Terminology collections and domain specific data should receive greater attention.
- High-quality basic LT tools, such as spell checkers, should be developed for lesser-used languages. On a wider scale, a common technological approach to building language tools must be found so that their design and development is made more efficient.
- Research into code-switching and "hybrid" language use, such as Hinglish, should be encouraged, particularly due to their prevalence in social media
- Research could be channeled by determining a set of convincing multilingual use cases that have wide appeal.

Technology and Data Recommendations

- The most frequent recommendations touched on the need for open-source data. Feedback pointed to the need to guarantee data and resources are made publicly accessible under public licences that ensure reuse, particularly when they have been developed using public funding. Intellectual property rights regulation must be more flexible and allow for greater utilization of protected data for the development of language technology. In some cases, specific domains should be targeted, such as health care. Similarly, open-source software could allow small- and medium-sized companies to develop applications in their preferred languages with an initial investment. Small-market languages would avoid dependence on proprietary solutions from large multinational companies, many of which are often reluctant to create high-end applications for them. Indeed, similar approaches that encourage data sharing are needed to increase opportunities for less-spoken languages. Furthermore, the need to ensure data is open-sourced should be couched in an overall call to foster an open-source culture.
- Solutions must be found to fill gaps in state-of-the-art natural language understanding and generation. Large and more complex datasets, as well as high-performance computing resources are needed. Several MRLUs require a wider range of tools, higher quality applications, and more and higher quality data. The same is true of large language models, which are scarce for many languages.

- It may be possible to utilize more sophisticated English tools and resources to continue development of cross-lingual transfer learning (CLTL) in order to build NLP models for low-resource target languages. This may be accomplished by leveraging labelled data or via a staged process whereby training data from English feeds the development of languages with moderate resources.
- More sophisticated English tools and resources may also be utilized to create multilingual transfer settings that enable training data in multiple source languages to be leveraged to further boost performance of low-resource languages.
- In a bilingual (majority / minority language) environment, bilingual models are needed to enable development of bilingual tools that will facilitate working in such environments. For instance, in the case of Welsh, English-Welsh models are needed.
- Broaden the scope of language varieties that NLP tools can handle by developing annotated resources and tools for non-standard varieties.
- A key recommendation involves strengthening the ties and cooperation between public administrations, academia and industry with respect to LT development and use. Whereas academia often reacts slowly to rapid change, industry generally mobilizes in only specific areas and does not engage in sufficient research.
- Resources other than text corpora are also needed for lesser-spoken languages. Multimodal resources for several languages are still unavailable and targeted actions are required to fill observed gaps in speech and other multimodal data. This includes creating more language resources and tools.
- Citizen science or crowd-sourcing are potential approaches to data collection, dataset creation and tool evaluation. In a related fashion, user-driven quality assessment of LT tools and services could prove beneficial. The same is true of access to basic LT functionality, such as voice tech and spell checking.
- More focus should be placed on adapting tools and resources to areas such as digital humanities which are resistant to adopting LT. Lowering the barrier towards tool usage is important: tools and applications that can be used by non-experts in LT are necessary.
- Specialized datasets with material for domain-specific purposes to adapt general-purpose language models are needed, as is language data from emerging domains, such as social media.
- Multilingual support for automatic annotation tools.

Infrastructure Recommendations

- Infrastructures and trained personnel are a must. Stable funding for maintenance of widely used, national-level language resources and infrastructures should be provided and the sustainability of existing resources should be addressed. Several experts suggested creating national and local centres or repositories dedicated to LT. These would centralize and standardize LT resources and tools. They could also facilitate dialogue between stakeholders.

Policy Recommendations

- Europe is lagging behind Asia and the US, but it should attempt to ensure its sovereignty in language technology. European languages need European technologies, European LT infrastructure and a thriving LT ecosystem that includes academia, industry and startups. To attain these, the EC should grab hold of its previous message concerning LT funding. By the same token, measures must be put in place that ensure LT and language-centric AI is appropriately recognized and included in state policies for language, cultural and technological development. A shift in focus is required to recognise technology as an equally important axis for continued language use.

- Feedback often focused on the need for long-term and dedicated funding programmes at the local, national, and European levels. Languages with fewer speakers are especially in need of such funding if the languages are to be protected in the digital age. This is partly because the LT market is unable to provide sufficient support for them. Funding should be directed to language-specific LT and especially to the particular needs of digitally endangered languages.

- The need for long-term and dedicated funding is tied to the recommendation that the public sector should continue to support the development of LT tools and resources through national and international LT programmes. Ideally, they will also encourage the collection, preparation and distribution of as much data as possible. These programmes should provide support for both research and industry, as well as for low-resource languages. The latter is important because there is a strong belief that LT tools and applications can not only significantly improve digital literacy for MRLU languages, but also grow the number of their speakers through online translation tools and other applications. In this way, meaningful language equality can be achieved.

- Increase the innovation capacity of public services. The best way to improve service to citizens and at the same time increase the demand for technology is to strengthen their capability for technological innovation.

- International IT companies must be convinced to include more European languages in their products.

- Another area of consensus among the feedback received focused on the need for LT and AI programmes in education. Such programmes would help address the general lack of skilled LT professionals. Schools and universities should be encouraged to teach NLP and computational linguistics within an interdisciplinary approach. Doing so would help further digital literacy and awareness about tools in research communities and society at large.

- Opportunities for cross-border cooperation on LT between governments should be explored.

5 Conclusion

The final round of feedback on the SRIA and LT for individual languages helped consolidate the ELE consortium strategic recommendations on how to advance LT and language-centric AI research, technology, infrastructure, and policy. Given Europe's varied linguistic landscape, a hallmark of its cultural heritage, it is unsurprising that perceived lacunae and proposals for specific languages differ from one to the next. Yet, despite these differences, born of the historical, political and socioeconomic peculiarities of each language, transversal solutions to strengthen LT applications and mitigate the digital disparities between languages are evident. Among the most pressing are the need for the creation of a coordinated effort in the form of interconnected regional, national and international LT plans, long-term and dedicated funding and open-source data. Underlying these goals is a strong belief that working towards Digital Language Equality must entail open, concerted, interdisciplinary and cross-sector engagement from all stakeholders. In this spirit, EU coordination of LT should be complemented by efforts at the national and regional levels. The ELE project has sought to deepen this sentiment of cooperation and prepare the road ahead to be paved with concrete actions. The recommendations gathered here will aid in doing so by designing and establishing a shared European programme for Language Technology and Digital Language Equality.

References

Gilles Adda, Annelies Braffort, Ioana Vasilescu, and François Yvon. Deliverable D1.14 Report on the French Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_14__Language_Report_French_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. Deliverable D1.2 Report on the State of the Art in Language Technology and Language-centric AI, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Itziar Aldabe, Georg Rehm, German Rigau, , and Andy Way. Deliverable D3.1 Report on existing strategic documents and projects in LT/AI, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Dimitra Anastasiou. Deliverable D1.24 Report on the Luxembourgish Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_24__Language_Report_Luxembourgish_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Gerhard Backfried, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz. Deliverable D2.14 Technology Deep Dive – Speech Technologies, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_14__Speech__Technologies.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

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. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Lars Borin, Rickard Domeij, Jens Edlund, and Markus Forsberg. Deliverable D1.33 Report on the Swedish Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/05/ELE___Deliverable_D1_33__Language_Report_Swedish_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

António Branco, Sara Grilo, and João Silva. Deliverable D1.28 Report on the Portuguese Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_28__Language_Report_Portuguese_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Aivars Bērziņš, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, and Andy Way. Deliverable D2.13 Technology Deep Dive – Machine Translation, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_13__Machine_Translation_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Kristine Eide, Andre Kåsen, and Ingerid Løyning Dal. Deliverable D1.26 Report on the Norwegian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_26__Language_Report_Norwegian_.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. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Anželika Gaidienė and Aurelija Tamulionienė. Deliverable D1.23 Report on the Lithuanian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_23__Language_Report_Lithuanian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Radovan Garabík. Deliverable D1.30 Report on the Slovak Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_30__Language_Report_Slovak_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

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. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Maria Gavriilidou, Maria Giagkou, Dora Loizidou, and Stelios Piperidis. Deliverable D1.17 Report on the Greek Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_17__Language_Report_Greek_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Maria Giagkou, Penny Labropoulou, Stelios Piperidis, Miltos Deligiannis, Athanasia Kolovou, and Leon Voukoutis. Deliverable D1.37 Report on Database and Dashboard, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/05/ELE___Deliverable_D1_37__Dashboard__compressed.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Jose Manuel Gomez-Perez, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, and Teresa Lynn. Deliverable D2.15 Technology Deep Dive – Text Analytics, Text and Data Mining, NLU, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_15__Text_Analytics_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

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. 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. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Stefanie Hegele, Barbara Heinisch, Antonia Popp, Katrin Marheinecke, Annette Rios, Dagmar Gromann, Martin Volk, and Georg Rehm. Deliverable D1.16 Report on the German Language, 2022a. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_16__Language_Report_German_.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, 2022b. URL 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. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Davyth Hicks. Deliverable D2.9 Report from ELEN, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_9__Report_from_ELEN_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Jaroslava Hlavacova. Deliverable D1.8 Report on the Czech Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_8__Language_Report_Czech_.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. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Kinga Jelencsik-Mátyus, Enikő Héja, Zsófia Varga, Tamás Váradi, László János Laki, and Győző Yang Zijian. Deliverable D1.18 Report on the Hungarian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_18__Language_Report_Hungarian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Martin Kaltenboeck, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg. Deliverable D2.16 Technology Deep Dive – Data, Language Resources, Knowledge Graphs, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_16__Data_and_Knowledge_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Sabine Kirchmeier. Deliverable D2.8 Report from EFNIL, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_8__Report_from_EFNIL_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Svetla Koeva and Valentina Stefanova. Deliverable D1.5 Report on the Bulgarian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_5__Language_Report_Bulgarian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Simon Krek. Deliverable D1.31 Report on the Slovenian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_31__Language_Report_Slovenian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Cvetana Krstev and Ranka Stanković. Deliverable D1.35 Report on the Serbian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_35__Language_Report_Serbian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Krister Lindén and Wilhelmina Dyster. Deliverable D1.13 Report on the Finnish Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_13__Language_Report_Finnish_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Teresa Lynn. Deliverable D1.20 Report on the Irish Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_20__Language_Report_Irish_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Bernardo Magnini, Alberto Lavelli, and Manuela Speranza. Deliverable D1.21 Report on the Italian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_21__Language_Report_Italian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Diana Maynard, Joanna Wright, Mark A. Greenwood, and Kalina Bontcheva. Deliverable D1.11 Report on the English Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_11__Language_Report_English_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Maite Melero, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas. Deliverable D1.6 Report on the Catalan Language, 2022a. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_6__Language_Report_Catalan_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Maite Melero, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas. Deliverable D1.32 Report on the Spanish Language, 2022b. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_32__Language_Report_Spanish_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Sjur Nørstebø Moshagen, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, and Per Langgård. Deliverable D1.38 Report on the Nordic Minority Languages, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/05/ELE___Deliverable_D1_38__Language_Reports_nordic_languages_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Kadri Muischnek. Deliverable D1.12 Report on the Estonian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_12__Language_Report_Estonian_.pdf. Reports on European Language Equality (ELE) | Coordinator: Prof. Dr. Andy Way, Co-Coordinator: Prof. Dr. Georg Rehm, received funding from the European Union (Grant Agreement no. LC-01641480 – 101018166 ELE).

Maciej Ogrodniczuk, Piotr Pęzik, Marek Łaziński, and Marcin Miłkowski. Deliverable D1.27 Report on the Polish Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_27__Language_Report_Polish_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Bolette Sandford Pedersen, Sussi Olsen, and Lina Henriksen. Deliverable D1.9 Report on the Danish Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_9__Language_Report_Danish_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Delyth Prys, Gareth Watkins, and Stefano Ghazzali. Deliverable D1.34 Report on the Welsh Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_34__Language_Report_Welsh_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Vasile Păiș and Dan Tufiș. Deliverable D1.29 Report on the Romanian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_29__Language_Report_Romanian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Mike Rosner and Claudia Borg. Deliverable D1.25 Report on the Maltese Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_25__Language_Report_Maltese_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

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.

Eiríkur Rögnvaldsson. Deliverable D1.19 Report on the Icelandic Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_19__Language_Report_Icelandic_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Kepa Sarasola, Itziar Aldabe, Arantza Diaz de Ilarraza, Ainara Estarrona, Aritz Farwell, Inma Hernaez, and Eva Navas. Deliverable D1.4 Report on the Basque Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_4__Language_Report_Basque_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Inguna Skadiņa, Ilze Auziņa, Baiba Valkovska, and Normunds Grūzītis. Deliverable D1.22 Report on the Latvian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_22__Language_Report_Latvian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Frieda Steurs, Vincent Vandeghinste, and Walter Daelemans. Deliverable D1.10 Report on the Dutch Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_10__Language_Report_Dutch_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

José Manuel Ramírez Sánchez and Carmen García Mateo. Deliverable D1.15 Report on the Galician Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_15__Language_Report_Galician_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Marko Tadić. Deliverable D1.7 Report on the Croatian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_7__Language_Report_Croatian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

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. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Andy Way, Georg Rehm, Jane Dunne, Maria Giagkou, José Manuel Gomez-Perez, Jan Hajič, Stefanie Hegele, Martin Kaltenböck, Teresa Lynn, Katrin Marheinecke, Natalia Resende, Inguna Skadiņa, Marcin Skowron, Tea Vojtěchová, and Annika Grützner-Zahn. Deliverable D2.18 Report on the state of Language Technology in 2030, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/05/ELE___Deliverable_D2_18__Report_on_State_of_LT_in_2030_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Tarik Ćušić. Deliverable D1.36 Report on the Bosnian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_36__Language_Report_Bosnian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Appendix

A Detailed Feedback

A.1 Feedback from Basque

- A significant gap remains between Basque and other languages in terms of data
- In comparison to text corpora, the amount of resources for Basque that include other modalities is relatively small
- To guarantee data and resources will be made publicly accessible and to ensure resources resulting from public funding are publicly available
- LT tools available for Basque can be used by administrations, institutions and companies to create at no great economic cost many more documents in Basque
- Infrastructures and trained personnel are required

A.2 Feedback from Bulgarian

- General matters for SRIA:
  - Large multimodal resources for Bulgarian are still unavailable
  - Large models for Bulgarian and sample-efficient pre-training settings are still scarce
  - Strategic and program documents, targeted funding and the transfer of good practices, are still insufficient
- Language specific recommendations:
  - A lot of dedicated funding is needed (national, regional, European)
  - Need for dedicated LT and AI programs in education
  - Strong support to the development of open-source culture and collaboration

A.3 Feedback from Catalan

- General matters for SRIA
  - Political commitment at the EU level as well as the Member States level to truly facilitate the path towards reutilisation of all kinds of language data (text documents, translation memories, audiovisual materials, etc..)
  - Public-funded European hub and repository for ready-to-use, datasets, models and open-source tools (ala Hugging Face)
- Language specific recommendations for Catalan
  - Support for open source solutions. Open data and open-source software allow small and medium-sized companies (and potentially even large ones) to develop applications in Catalan without having to face the initial investment barrier. These solutions also guarantee technological sovereignty in the face of dependence on proprietary solutions from large multinationals, who are not always willing to develop the most high-end applications for small-market languages.
  - Increase the innovation capacity of Catalan public services. Public administrations own internal consumption should act as a driver of demand. The best way to improve service to citizens and at the same time increase the demand for technology in Catalan is to increase their capacity for technological innovation.
  - Creation of an independent Centre dedicated to Language Technologies in Catalan. Said Center would provide sustainability to the language infrastructures and resources for Catalan generated by occasional investments such as the AINA project, as well as bring technology in Catalan to the market, by facilitating a dialogue between stakeholders (administration, research, language institutions, local industry, GAFAM).

A.4 Feedback from Dutch

- Dutch is in a good shape but public corpora become quickly outdated (often made before 2011)
- Language use in recently emerged domains needs to be collected, such as social media
- Setup of a new bi-national program for cooperation between Dutch and Flemish governments for construction of corpora that document recent language
  - Written
  - Spoken
  - microblog

A.5 Feedback from English

- Focus on research into code-switching and "hybrid" language use such as Hinglish, Pidgins etc because it's common especially in social media
- Focus on adapting tools and resources to areas such as digital humanities which are still quite slow / resistant to adopting LT
- Making use of the sophistication of English tools and resources by:

  - Continued development of cross-lingual transfer learning (CLTL) in order to build NLP models for a low resource target language by leveraging labelled data from languages such as English with a high level of resources, or via a staged process whereby training data from English feeds the development of languages with moderate resources;
  - Multilingual transfer settings enabling training data in multiple source languages to be leveraged to further boost performance of low-resource languages.

A.6 Feedback from Estonian

- General SRIA recommendations
  - Encourage publishing and effortless sharing of language data
  - More NLP courses for students, also make NLP as a career option more visible
  - Encourage collaboration between academia and industry
- Language-Specific Recommendations
  - Continue the creation of missing tools and resources, made available under public licences
  - Public sector should continue to support developing LT tools and resources
  - Take care that data-sharing is encouraged, all developed tools and resources are publicly available and the local NLP community knows where to look for them
  - Broaden the scope of varieties of Estonian that NLP tools can handle: develop annotated resources and tools for non-standard varieties of Estonian
  - Encourage universities to teach and students to study NLP and Computational Linguistics

A.7 Feedback from Finnish

- General purpose SRIA:
  - General-purpose language-centric AI models trained on cross-language and cross-domain resources, which can benefit from adaptation to local language varieties and specialized domains with small or medium-sized data sets.
- Language specific considerations:
  - A variety of specialized data sets with language materials for domain-specific purposes to adapt general purpose language models.

A.8 Feedback from Galician

- General matters for SRIA
  - A substantial effort to create LT resources that increase the diversity of domains and the number of multimedia resources.
  - Policies and regulations that guarantee open access to ready-to-use datasets based on public data of European, National, Regional and Local public institutions.
  - Put in place of initiatives to create multi-domain databases and resources for low-resourced languages.

- Language specific recommendations
  - Promote LT for Galician at the level of other co-official languages of Spain, such as Catalan or Basque.
  - Creation of a centre to centralize and standardize all the LT resources and tools created for Galician.
  - Increase the use of LT in Galician public services and institutions.

A.9 Feedback from Greek

- SRIA
  - maintenance, extension and sustainability of LT-related infrastructures
  - national and European coordinated actions for ensuring access to open high-performance computing infrastructures
  - coordinated actions for the development of large-scale LRs ready to power large language models supporting a wide range of applications
  - coordinated actions to promote the culture of data sharing, including open-source software, involving all stakeholders, the public sector, research and industry
- Language specific
  - targeted actions to fill-in the observed gaps in speech and multimodal data
  - measures ensuring that LT and language-centric AI is appropriately recognized and included in the state policies for language, cultural and technological development
  - actions to further enhance digital literacy in the research communities and the society as a whole

A.10 Feedback from Hungarian

- General matters for SRIA
  - the size of existing raw corpora still needs to be increased - especially to train language models
  - more annotated benchmark corpora should be compiled for evaluation purposes
  - regulation for the access and re-use of language data is missing
- Language specific recommendations
  - still room for cooperation between research and industry/public administration
  - good quality and well organised LT education needed
  - for lesser used languages like Hungarian a lot of national/local funding is needed - as for these languages LT-connected market in itself is unable to provide sufficient financial background

A.11 Feedback from Icelandic

- General matters for SRIA
  - Low-resourced languages where the market is small and not sustainable need special support.

- Language specific recommendations
  - Continued funding for the National Language Technology Programme must be ensured
  - Major international IT companies must be convinced to include Icelandic in their products
  - Cooperation between the industry and academia must be strengthened

A.12 Feedback from Irish

- General matters for SRIA
  - Improvements in access and re-use of language data
  - Additional support needed for low-resourced languages where technology is not market-driven (ie no investment from industry)
- Language specific recommendations
  - Change of Focus To date, the Irish language has received much investment into the development of dictionaries and terminologies due to a primary focus on supporting translators and Irish language learning. However, a shift in focus is required to recognise technology as an equally important axis for continued language use.
  - Need for Dedicated LT Programmes/ Skill shortages It is particularly difficult to source experts with the right combination of skills (e. g. Irish language, computer science, linguistics) to further LT research for Irish. Currently only one university in Ireland offers this type of inter-disciplinary course at undergraduate level.
  - Funding/ Long-term strategy In the absence of a Digital Language Strategy, as yet, there are no long term funding schemes or research centres dedicated to Irish LT. This needs to change to ensure a strategic plan for safeguarding Irish in a digital age.
  - Untapped Potential As the value of language data is broadly unknown amongst Irish citizens and across the Irish public sector, there is much untapped yet currently inaccessible data that could make a huge impact on the future of Irish LT. Also, the general positive disposition and altruistic nature of Irish speakers toward supporting the language should be leveraged more through citizen science or crowd-sourcing approaches to data collection, dataset creation and tool evaluation.

A.13 Feedback from Italian

- General matters for SRIA
  - Need of more freely available documents to train models
- Language specific recommendations
  - Promote the production of freely available general-purpose resources
  - Remedy to poor availability of domain-specific resources
  - Promote collaboration between academia and industry

A.14 Feedback from Latvian

- General matters for SRIA
  - Less resourced languages need special support
  - Close synchronization between national and international activities is necessary, especially, with respect to research infrastructures and research priorities
- Latvian language specific recommendations:
  - Need for dedicated long-term LT programs that provide equal support for both research and industrial activities
  - Strong support for the creation of missing resources, support for open access LRTs, opening LRTs from public sector and public funded projects. There are still significant gaps with respect to solutions that involve deep state of the art natural language understanding and generation, require large and complicated datasets and high performance computing resources.
  - Provide stable funding for maintenance of widely used, national level language resources and infrastructures

A.15 Feedback from Lithuanian

- General matters for SRIA
  - Support for low-resourced languages.
- Language specific recommendations
  - Open Access Language Resource Infrastructure. Lithuania still lags behind in data sharing culture, with unresolved licensing issues - intellectual property rights regulations, which need to be more flexible and allow for greater use of protected data for language technology development and resources.
  - Human Resources. Lithuania lacks the necessary human resources: there is a lack of IT specialists in language technologies, as well as researchers in this field, and there are no specialized study programs.
  - National and international support. Lithuania needs support, including dedicated long-term language technology programs; cooperation between research and industry / public administration is needed in this area.

A.16 Feedback from Maltese

- General matters for SRIA
  - Emphasis on the crucial role of LT within the big issues of intelligent interaction, knowledge management, trust, conflict resolution.
- Language specific recommendations
  - Increased involvement of industry, particularly the thriving IT industry, in LT use and development.
  - Coming up with a set of convincing multilingual use cases that will have wide appeal, not necessarily involving the language pair Maltese/English
  - Access to basic BLARK-style functionality e.g voice tech and spell checking.
  - LT for health and other specific domains.
  - Better management of the resources created locally (links, centralised repository,...)

A.17 Feedback from Norwegian

- General SRIA recommendations
  - Develop standards for user-driven quality assessment
  - Continue to raise awareness of the importance of language data.
- Language specific recommendations
  - Continue the creation of missing tools and resources, made available under permissive licences to ensure their reusability.
  - Ensure sufficient funding for language-specific LT for Bokmål and Nynorsk.
  - Public sectors must take on their new responsibility as required in the new language act and ensure parallel versions of Norwegian Bokmål and Norwegian Nynorsk language technology in public procurement.
  - Downstream (user-driven) quality assessment of Norwegian language technology tools and services in order to compare the quality of Nynorsk and Bokmål tools and services as well as dialect understanding.

A.18 Feedback from Polish

- What we need the most:
  - HPC for European science to boost AI/NLP development,
  - leading to development of standard optimized language models
- Language-specific recommendations:
  - Provide support for Ukrainian LRTs
  - Provide stable funding for maintenance of crucial language resources for Polish such as The National Corpus of Polish or The Great Dictionary of Polish (but also for building standard domain-specific resources such as the Polish SNOMED)
  - Increase support for low-resources languages in Poland: even the largest ones (Kashubian and Silesian) are not adequately supported

A.19 Feedback from Portuguese

- What really matters for the SRIA?
  - to pass a key message:
    - LT is a key geopolitical asset and even more so for EU, where multilingualism is an vital challenge
    - EC needs to recover its previous informed vision concerning LT, with ambitious specific funding support for Research, Development and Innovation for LT
- Main recommendations for Portuguese?
  - Ambitious dedicated national programme for the technological preparation of the Portuguese Language for the digital language.
  - That includes gathering, preparing and openly distributing as much language and multimodal data as possible, raw and labelled.
  - And includes promoting research on the Portuguese language

A.20 Feedback from Serbian

- Language specific recommendations
  - A lot of dedicated and long-term funding is needed (national, regional, European), especially if open resource and LT tools and applications are needed; this can foster stronger collaboration between research and industry/public administration.
  - Need for dedicated LT and AI programs in education that would bring together language and computational specialists.
  - Establishment of a center dedicated to the production and promotion of language resources and technologies for Serbian

A.21 Feedback from Slovak

- General SRIA
  - allow re-use of available data without clear licensing
  - address fragmentation of resources
- Language specific recommendations
  - LT in research and industry/public administration are quite divorced:
    - ∗ academy is often reacting very slowly
    - ∗ industry is mostly interested in very specific areas (& does not do research)
    - ⇒ aim for closer collaboration
  - clarify (open) licensing for many existing datasets
  - many resources remain in “proof of concept” stage, effort is needed to bring them up to usability
  - address sustainability of existing resources

A.22 Feedback from Spanish

- Language specific recommendations for Spanish
  - Well-regulated access to linguistic data. A convenient and well-regulated access to data is essential for the development of new products, applications and services. Appropriate open data policies based on ethics, transparency and accessibility to data from both the private and public sector (including administrations, the public broadcasting corporation, etc.) must be promoted, while guaranteeing citizens' rights to privacy and confidentiality.
  - Increase the innovation capacity of Spain's public services. Public administrations own internal consumption should act as a driver of demand. The best way to improve service to citizens while at the same time stimulating the LT market and industry in Spanish is to increase their capacity for technological innovation.
  - Creation of an independent Centre dedicated to Language Technologies in Spanish. Said Center would provide sustainability to the language infrastructures and resources generated by the Plan of Impulse of LTs, create synergies between stakeholders (administration, research, language institutions, local industry, GAFAM), as well as be a reference point for entities dealing with other languages of Spain, and other Spanish speaking markets.

A.23 Feedback from Welsh

- SRIA
  - Develop methods for legislation which facilitate collection of open data together with clarifying the legal position on re-use of data in Europe
  - Provide their own funding programme(s) for minoritized / endangered languages to develop basic language tools and resources to ensure their digital survival
  - Develop exchange programmes to share knowledge including specific programmes to up-resource and up-skill minoritized / endangered languages
  - Enable minoritized / endangered languages to join in large-scale multinational and multilingual R&D programmes of the type previously reserved for official EU languages
- To fill the gaps for LT provision for Welsh
  - Bilingual (English / Welsh) models to enable development of bilingual tools to facilitate working in a bilingual (majority / minority language) environment
  - Programme to maintain and further develop existing tools and resources as well as funding new projects

A.24 Feedback from ECSPM

- General recommendations
  - The situation for the smaller official languages of the EU and those considered community or heritage languages in the EU is dire, but the situation for Minority/Regional/Lesser-Used Languages (MRLUs) is even worse – a situation which cannot be improved without EU and national funding, allocated especially for the digital support of the MRLUs.
  - Both on national and EU level, it is important to raise awareness among politicians,, for need of the digital support of lesser used languages.
  - It is important to create opportunities for people who are speakers of the MRLUs to study in areas related to LT.
- For the languages we were responsible for
  - A wider range of tools and applications of higher quality are needed for heritage or community languages and for the MRLUs
  - Our informants are convinced that LT tools and applications will significantly improve literacy in the MRLUs and increase the number of speakers, so it is important that they be supported to develop translation tools, applications and materials online
  - In further investigating the MRLUs digital support needs, there should not be exclusive use of English. Many of our respondents had low proficiency in the language.

A.25 Feedback from EFNIL

- Development and promotion of basic tools (spell checkers, speech tools search and MT) should have better quality and be the first priority for lesser used languages.
- General awareness of the existence and functionality of tools should be increased at the user side (language technology and AI as school subjects).

- Terminology collections and domain specific data should have more attention.

A.26 Feedback from ELEN

- There needs to be dedicated funding for RML development separate from that of the EU official languages and that research, funding and support is needed on the specific needs of each RML/ endangered language. Work on the ELE Project illustrated the huge disparities between RMLs in terms of LT development and the difficulties in obtaining data from some endangered language communities.
- Input from our members focuses on the need for the ELE project to treat all European languages equally instead of focusing on the EU official languages and a few co-official ones. There is a call for written reports for each language, and the finance for that, just as there has been for each EU official language, and that this is not as some kind of voluntary extra. For example, members asked why there was a separately financed report for Icelandic (not an EU language) but not one for Frisian or Gaelic (both co-official)?
- The results from the survey have been very useful in illustrating the critical lack of LT provision for RMLs and endangered languages. For the future our members would like to see more emphasis and resources put on these languages so that we can achieve meaningful language equality and levelling up, and so that none of these languages are left behind.

A.27 Feedback from Eurescom

- For minority and less spoken languages, in respect to all kind of language tools (recognition, support, learning, etc.), we need to find a (technological) way to consider it within a common approach, in order to create synergies and increase efficiency of the solutions and their design and development
  - is it possible to leave "single-language" approach?
  - would it mean concentration of the available funding Europe Wide?
- Open source and further similar approaches
  - To increase opportunities for minority/less spoken languages
  - to increase overall language support, reduce costs, and ensure sustainability of the solutions

A.28 Feedback from Wikimedia

- What really matters for the Strategic Research and Innovation Agenda?
  - Availability of open-source material - especially for minority/regional/lesser used languages (language learning materials, school books, open-source dictionaries, translations resources, stop words, stemmers, written documents, audio data or spell checkers),
  - Translation Tools for minority/regional/lesser used languages (real-time and collaborative translations tool in multiple languages; translating text and documents to and from multiple languages),
  - More people contributing to their minority/regional/lesser used languages languages online (e.g. Wikipedia articles, Wikidata Lexemes),

  - Lack of long-term funding for projects and institutions (e. g., libraries) working with regional and minority languages

A.29 Feedback from Ontotext

- What really matters for the Strategic Research and Innovation Agenda?
  - Multilingual support and open access availability of resources: vocabularies, taxonomies, ontologies, domain specific terminologies
  - Standardization:
    - Lack of multilingual benchmark datasets for evaluation of language technologies
    - Lack of standardized workflows for annotated corpora generation and support
  - Multilingual support of the automatic annotation tools

A.30 Feedback from SAP

- More high-quality linguistic training data that is relevant, balanced, available under reasonable licensing terms, and adheres to the FAIR data principles (e.g., utilising Linked Data approaches)
- More diverse, recent, and properly annotated evaluation data as well as corresponding evaluation environments that anyone can use (e.g., to evaluate a speech recognition with own audio)
- Methods and capabilities that consider the interactional, and communicative contexts of language (e.g., take speech acts into account)
- Algorithms that include explainability, guaranteed performance, knowledge ingestion and extensibility, as well as easy error correction (e.g., without extensive retraining)
- Data usage rights that facilitate NLP research and development (e.g., see existing GDPR exceptions for research (e.g., in medicine))
- Investment protection and interoperability driven by official standards (e.g., for provenance from the World Wide Web consortium), or industry-standards (e.g., for entity types from schema.org)
- Requirements/conformance statements for Language Technology artefacts; in the context of regulated industries, certification – the assignment of a label based on transparent testing, and compliance with conformance criteria – may need to be considered
- Consumer-grade tool support for domain experts that allow them to generate, label, access and process structured data/knowledge (e.g., knowledge graphs), or to generate, use and evaluate language processing results (e.g., recall and precision of term linking solutions)