The European Language Equality Project: Designing a Roadmap for Digital Language Equality

Itziar Aldabe[1], Aritz Farwell[1], Eva Navas[2], Inma Hernaez[2], German Rigau[1]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]HiTZ Center - Aholab, University of the Basque Country UPV/EHU
{itziar.aldabe,aritz.farwell,eva.navas,inma.hernaez,german.rigau}@ehu.eus

Abstract

This paper provides an overview of the European Language Equality (ELE) project. The main objective of ELE is to prepare the European Language Equality program in the form of a strategic research and innovation agenda that may be utilized as a road map for achieving full digital language equality in Europe by 2030. The desk research phase of ELE concentrated on the systematic collection and analysis of the existing international, national, and regional strategic research agendas, studies, reports, and initiatives related to language technology and LT-related artificial intelligence. A brief survey of the findings is presented here, with a special focus on the Spanish ecosystem.

Index Terms: Digital Language Equality, Strategic research and innovation agenda, Desk research

1. Introduction

Natural language is at the heart of human intelligence. It is the most common and versatile way for humans to convey and access information. We use languages to encode, store, transmit, share and manipulate it. Perhaps unsurprisingly, most available digital information is unstructured in the form of written, spoken, or multimodal documents in multiple languages. In fact, up to 80% of all data is unstructured [1], presenting a challenge for any organization that wishes to exploit and process it utilizing computer systems. This is because it is non-trivial and intrinsically complex to process *unstructured digital information*, including written and spoken language, which is subject to multiple interpretations (ambiguity) and requires world knowledge and also contextual knowledge about specific situations. Partly for these reasons, language is central to our efforts to develop practical AI technologies [2].No sophisticated and effective AI-powered tool can exist without mastery of language, making it the next great frontier in AI [3, 4, 5].

With this in mind, it is easy to understand why achieving digital linguistic equality in a linguistically rich geographical area is a consequential and challenging endeavor on many fronts. Europe, home to about 800 million individuals spread across 10 million km^2, presents one such case. Twenty-four official languages and more than 60 regional and minority languages constitute the fabric of the EU's linguistic landscape. Moreover, multilingualism is a cultural cornerstone of Europe and signifies what it means to be and feel European. Unfortunately, language barriers still hamper communication and the free flow of information across the European Union. And while language technologies offer a means to mitigate these obstacles, many studies have found a striking imbalance between languages in terms of support through LT.

In 2018, the European Parliament resolution "Language equality in the digital age" issued a call to remedy this situation. Starting in January 2021, the European Language Equality (ELE) project [6] answered this call and laid the foundations for a strategic research, innovation and implementation agenda (SRIA) and road map to make full digital language equality (DLE) a reality in Europe by 2030. Developing the SRIA and road map involved many stakeholders with different perspectives. Accordingly, the ELE project – led by DCU, with DFKI, Charles University, ILSP and EHU/UPV as core members – put together a large consortium of 52 partners, who together with the wider European LT community, prepared the SRIA and roadmap for all European languages.

The following sections summarize the desk research that was undertaken in ELE, which involved the systematic collection and analysis of the existing international, national and regional strategic research agendas (SRAs), studies, reports and initiatives related to language technology and LT-related artificial intelligence. Around 200 such documents from both within and without the European Union were reviewed for this purpose.

2. National Language Technology Initiatives in Europe

In 2018, as part of the Spanish *Plan for the Advancement of Language Technology*, an analysis of the LT landscape in Europe was presented [7]. European strategies, policies and programmes to support LT were identified. At the national level, they concluded that there are or have been only very few dedicated national programmes designed to finance projects related to LT. In contrast, in most countries funding for the development of LT is provided through generic R&D calls. At the European level, so far LT has received better support through calls in different programmes: FP7, H2020, CEF Telecom, CIP ICT-PSP, EUREKA and EUROSTARS, among others. However, the authors note that in the most recent programmes funding for LT projects has been gradually reduced.

Rehm, et al.[8] present an overview of various European LT and AI reports. As part of the European Language Grid, the 32 ELG National Competence Centres contributed information about the national funding situation for AI- and LT-related topics in their respective countries [9, 10, 11].

If we compare both studies, we can observe a small increase in the number of language-centric AI initiatives in the last couple of years. The effort of all European countries to be in line with the EC, which considers AI an area of strategic importance is noteworthy. In December 2018, the EC and the Member States published the "Coordinated Plan on Artificial Intelligence" COM(2018)795 [12], on the development of AI in the EU. The number of EU countries with an AI strategy (29 out of 30, 97%) demonstrates the success of the plan. Only Croatia has

no official AI strategy as of yet. Figure 1 presents an overview of the LT funding situation in Europe. In green those countries with a dedicated LT programme. In orange those countries that explicitly provide funding for LT-related topics through AI. In yellow the ones with an AI strategy mentioning LT and finally, in red, those countries with no AI strategy or having an AI strategy but not mentioning LT at all.
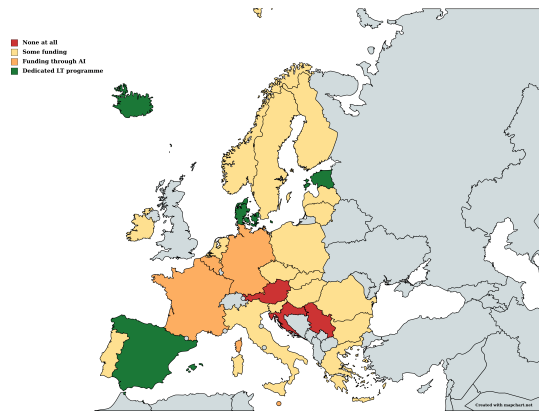

Figure 1: *Overview of the LT funding situation in Europe*

The report [13] also analyses the EU national AI strategies to identify areas for synergies and collaboration. It identifies several policy areas: human capital, from lab to market, networking, infrastructure, regulation. LT is mentioned as part of the Danish, Latvian, Maltese, Portuguese, Slovakian, Spanish and Swedish initiatives, so, as [8] mention, LT is finally generating some momentum. LT is considered key, and language understanding is seen as one of the next generations of innovative AI technologies [14]. For that, it is indispensable to set aside funding exclusively for LT. According to [8], only four of the 30 surveyed countries do not have some type of LT funding. Four countries have programmes dedicated to LT (in green in Figure 1, Denmark, Estonia, Iceland, Spain), six provide funding for LT-related topics through AI (in orange in Figure 1, Belgium, Denmark, Estonia, France, Germany, Malta) and two (Ireland, Latvia) that do not have LT programmes, but rather a language strategy defined by their governments.

The Spanish government has recently announced a new strategic project for economic recovery and transformation (PERTE in Spanish) called "The New Language Economy" [15]. The PERTE is presented as an opportunity to take advantage of the potential of Spanish and co-official languages as a factor for economic growth and international competitiveness in areas such as AI, translation, learning, cultural dissemination, audiovisual production, research and science. To do this, it has a budget of 1.1 billion euros of public investment, with the aim of mobilizing another billion in private investment. Moreover, following the lines of the Spanish Plan for the Advancement of LT [16], the Catalan government has also launched the AINA project [17], the Galician government the Nós project [18] and the Basque Country the GAITU project [19].

In summary, there are or have been only very few dedicated nationally financed programmes related to LT. In contrast, national funding for the development of LT is provided through generic R&D calls, not specific programmes. Moreover, only 12 European countries out of the 30 studied explicitly consider LT within their national policy initiatives. There is an immense need for pan-European coordination for putting LT into the na-tional AI strategies and into the policies for language, cultural, and technological development.

3. Main Recommendations

In line with the final recommendations of the EP *Language equality in the digital age* resolution [20], the main recommendations put forward by the reports, documents and initiatives analysed in this project emphasises the need of an European leadership in LT by creating a specific programme tailored to Europe's needs and demands i. e., to establish a large-scale, long-term coordinated funding programme for research, development, innovation and education in the field of LT, at European, national and regional levels. This programme has two main goals:

- Main societal and economic goal: digital language equality in Europe in 2030

- Main scientific goal: deep natural language understanding in 2030

In addition, it is necessary to allocate the area of *multilingualism and language technology* to the portfolio of a Commissioner, within the EC and to ensure comprehensive EU-level legal protection for the more than 60 regional and minority languages. To ensure the development of LT for official EU languages which are less widely spoken by coordinating the development of comprehensive language-related policies and developing strategies and policy actions to facilitate multilingualism in the digital market as well as to establish mechanisms to achieve European LT sovereignty by means of:

- Sufficient public sector data, data from broadcasters, social media, publishers, etc.

- Flexible access and support to sufficient GPU-based HPC facilities

- Sufficient LT experts in EU research centers and companies

Similarly, the EU should promote:

- that the Open Data Directive (2019/1024/EU) should include language data as a high-value data category and protocols for language data sharing are needed.

- a distributed centre for linguistic diversity to strengthen the awareness of the importance of lesser-used, regional and minority languages.

- a pan-European network of research centres.

- that all EC-funded projects have a language diversity plan.

- the development of better benchmarks and datasets for all languages, domains, tasks and modalities

- To extend LT beyond language

- To focus on language and culture-specific technologies (not just transfer from English)

- To enforce open ecosystems, open source, open access, open standards and interoperability

- To reinforce existing (and future) RIs (e.g. CLARIN, ELG, LDS, etc.)

It is also crucial to retain and attract talent in Europe. For that it is important to aim European education policies with this particular purpose and to create the necessary appealing conditions. Similarly, the development of fair and inclusive research and innovation requires the promotion of increased participation of women and representatives of ethnic and social minorities in the field and in order to ensure trust in LT, a regulatory framework for the development, implementation and deployment of reliable LT technologies is required.

The European SMEs and startups can be enable and empower by developing actions and appropriate funding as well as to secure a European leadership position in the field of language-centric AI since EU companies are the best placed to provide solutions tailored to our specific cultural, societal and economic needs. Finally, we have to engage administrations at all levels, and to use existing free and open-source LT, in order to improve the accessibility of those services.

These are the preliminary recommendations based on the initial desk research phase towards the systematic collection and analysis of the existing international, national and regional Strategic Research Agendas (SRAs), studies, reports and initiatives related to Language Technology (LT) and Artificial Intelligence (AI). These preliminary recommendations will be revisited, revised and extended, culminating in our final recommendations in the form of the strategic agenda and roadmap.

4. Acknowledgements

5. References

[1] "Search and unstructured data analytics: 5 trends to watch in 2020," https://www.accenture.com/us-en/blogs/search-and-content-analytics-blog/search-unstructured-data-analytics-trends, accessed: 2022-10-20.

[2] "The Power of Natural Language Processing," https://hbr.org/2022/04/the-power-of-natural-language-processing, accessed: 2022-10-20.

[3] "A.I. Is Mastering Language. Should We Trust What It Says?" https://www.nytimes.com/2022/04/15/magazine/ai-language.html, accessed: 2022-10-20.

[4] "Language Is The Next Great Frontier In AI," https://www.forbes.com/sites/robtoews/2022/02/13/language-is-the-next-great-frontier-in-ai/?sh=466c82315c50, accessed: 2022-10-20.

[5] "Is NLP innovating faster than other domains of AI," https://analyticsindiamag.com/is-nlp-innovating-faster-than-other-domains-of-ai/, accessed: 2022-10-20.

[6] "European Language Equality," https://european-language-equality.eu/, accessed: 2022-10-20.

[7] "Tecnologías del lenguaje en Europa," https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/tecnologias-del-lenguaje-en-Europa.aspx, accessed: 2022-10-20.

[8] G. Rehm, K. Marheinecke, S. Hegele, S. Piperidis, K. Bontcheva, J. Hajič, K. Choukri, A. Vasiļjevs, G. Backfried, C. Prinz, J. M. Gómez-Pérez, L. Meertens, P. Lukowicz, J. van Genabith, A. Lösch, P. Slusallek, M. Irgens, P. Gatellier, J. Köhler, L. Le Bars, D. Anastasiou, A. Auksoriūtė, N. Bel, A. Branco, G. Budin, W. Daelemans, K. De Smedt, R. Garabík, M. Gavriilidou, D. Gromann, S. Koeva, S. Krek, C. Krstev, K. Lindén, B. Magnini, J. Odijk, M. Ogrodniczuk, E. Rögnvaldsson, M. Rosner, B. Pedersen, I. Skadiņa, M. Tadić, D. Tufiș, T. Váradi, K. Vider, A. Way, and F. Yvon, "The European language technology landscape in 2020: Language-centric and human-centric AI for cross-cultural communication in multilingual Europe," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, pp. 3322–3332. [Online]. Available: https://aclanthology.org/2020.lrec-1.407

[9] "One Platform for all European Language Technologies," https://www.european-language-grid.eu, accessed: 2022-10-20.

[10] G. Rehm, M. Berger, E. Elsholz, S. Hegele, F. Kintzel, K. Marheinecke, S. Piperidis, M. Deligiannis, D. Galanis, K. Gkirtzou, P. Labropoulou, K. Bontcheva, D. Jones, I. Roberts, J. Hajič, J. Hamrlová, L. Kačena, K. Choukri, V. Arranz, A. Vasiļjevs, O. Anvari, A. Lagzdiņš, J. Meļņika, G. Backfried, E. Dikici, M. Janosik, K. Prinz, C. Prinz, S. Stampler, D. Thomas-Aniola, J. M. Gómez-Pérez, A. Garcia Silva, C. Berrío, U. Germann, S. Renals, and O. Klejch, "European language grid: An overview," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, pp. 3366–3380. [Online]. Available: https://aclanthology.org/2020.lrec-1.413

[11] G. Rehm, S. Piperidis, K. Bontcheva, J. Hajic, V. Arranz, A. Vasiļjevs, G. Backfried, J. M. Gomez-Perez, U. Germann, R. Calizzano, N. Feldhus, S. Hegele, F. Kintzel, K. Marheinecke, J. Moreno-Schneider, D. Galanis, P. Labropoulou, M. Deligiannis, K. Gkirtzou, A. Kolovou, D. Gkoumas, L. Voukoutis, I. Roberts, J. Hamrlova, D. Varis, L. Kacena, K. Choukri, V. Mapelli, M. Rigault, J. Melnika, M. Janosik, K. Prinz, A. Garcia-Silva, C. Berrio, O. Klejch, and S. Renals, "European language grid: A joint platform for the European language technology community," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 2021, pp. 221–230. [Online]. Available: https://aclanthology.org/2021.eacl-demos.26

[12] "Coordinated Plan on Artificial Intelligence," https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence, accessed: 2022-10-20.

[13] "AI Watch - National strategies on Artificial Intelligence: A European perspective, 2021 edition," https://ai-watch.ec.europa.eu/publications/ai-watch-national-strategies-artificial-intelligence-european-perspective-2021-edition_en, accessed: 2022-10-20.

[14] STOA, "Language equality in the digital age – Towards a Human Language Project," STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, 2017, http://www.europarl.europa.eu/stoa/.

[15] "PERTE Nueva economía de la lengua," https://planderecuperacion.gob.es/como-acceder-a-los-fondos/pertes/perte-nueva-economia-de-la-lengua, accessed: 2022-10-20.

[16] "Plan de Impulso de las Tecnologías del Lenguaje," https://plantl.mineco.gob.es/Paginas/index.aspx, accessed: 2022-10-20.

[17] "La nostra llengua es la teva veu," https://www.projecteaina.cat/, accessed: 2022-10-20.

[18] "Xunta e USC poñen en marcha o 'Proxecto Nós', que permitirá incorporar o galego ás novas linguaxes dixitais," https://www.xunta.gal/hemeroteca/-/nova/134792/xunta-usc-ponen-marcha-lsquo-proxecto-nosrsquo-que-permitira-incorporar-galego, accessed: 2022-10-20.

[19] "El Gobierno Vasco ha presentado GAITU, el Plan de Acción de las Tecnologías de la Lengua 2021-2024, el cual tiene el objetivo de integrar el euskera en las tecnologías lingüísticas," https://www.irekia.euskadi.eus/es/news/76846-gobierno-vasco-presentado-gaitu-plan-accion-las-tecnologias-lengua-2021-2024-cual-tiene-objetivo-integrar-euskera-las-tecnologias-linguisticas, accessed: 2022-10-20.

[20] European Parliament, "Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI))," http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf, 2018.