Overview of the ELE Project

Itziar Aldabe,[4] Jane Dunne,[1] Aritz Farwell,[4] Owen Gallagher,[1] Federico Gaspari,[1] Maria Giagkou,[5] Jan Hajic,[3] Jens Peter Kückens,[2] Teresa Lynn,[1] Georg Rehm,[2] German Rigau,[4] Katrin Marheinecke,[2] Stelios Piperidis,[5] Natalia Resende,[1] Tea Vojtěchová,[3] Andy Way[1]

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
[2] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
[3] Charles University (CUNI), Ovocný trh 5, Prague 1, 116 36, Czech Republic
[4] Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country) UPV/EHU, Barrio Sarriena s/n, 48940 Leioa, Bizkaia
[5]Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis (ILSP), Artemidos 6 & Epidavrou, GR-151 25 Maroussi, Athens, Greece

Abstract

This paper presents the ongoing European Language Equality (ELE) project, an 18-month action funded by the European Commission. The primary goal of the ELE project is to prepare the ELE programme, in the form of a strategic research, innovation and implementation agenda and roadmap for achieving full digital language equality in Europe by 2030.

1. Background

Twenty-four official languages and more than 60 regional and minority languages constitute the fabric of the EU's linguistic landscape. However, language barriers still hamper communication and the free flow of information across the EU. Multilingualism is a key cultural cornerstone of Europe and signifies what it means to be and to feel European. The landmark 2018 European Parliament resolution "Language equality in the digital age" found a striking imbalance in terms of support through language technologies (LTs) so issued a call to action. Starting in January 2021, ELE answered this call and is laying the foundations for a strategic research, innovation and implementation agenda (SRIA) and roadmap to make full digital language equality (DLE) a reality in Europe by 2030.

Developing an SRIA and roadmap for achieving full DLE in Europe by 2030 involves many stakeholders with different perspectives. Accordingly, the ELE project – led by DCU, and with DFKI, Charles University, ILSP and EHU/UPV as core members – has put together a large consortium of all 52 partners, who together with the wider European LT community, are preparing the different parts of the SRIA and roadmap, for all European languages: official, regional and minority languages.

2. Achievements & Ongoing Activities

Ensuring appropriate technology support for all European languages will create jobs, growth and opportunities in the digital single market. Equally crucial, overcoming language barriers in the digital environment is essential for an inclusive society and for providing unity in diversity for many years to come.

To date, we have concentrated on two distinct aspects: (i) collecting the current state of play (2021/2022) of LT support for the more than 70 languages under investigation, largely by the 32 National Competence Centres in our sister project European Language Grid (ELG);[2] and (ii) strategic and technological forecasting, i.e. estimating and envisioning the future situation in 2030 and beyond. Furthermore, we distinguish between two main stakeholder groups: LT developers (industry and research) and LT users as well as consumers. Both groups are represented in ELE by several networks (e.g. EFNIL, ELEN,

[2] https://www.european-language-grid.eu/

ECSPM) and associations (e.g. ELDA, LIBER) who each produce a report highlighting their own individual requirements towards DLE. The project's industry partners produce four "deep dives" with the needs, wishes and visions of the European LT industry regarding machine translation, speech technology, text analytics as well as data, all available on the project website. We have also organised a larger number of surveys and consultations with stakeholders who are not represented in the consortium.

We have formulated a preliminary working definition of DLE to drive our activities, namely: "*Digital Language Equality is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age."*

This DLE definition allows us to compute an easy-to-interpret metric (a "DLE score") for individual languages, which enables the quantification of the level of technological support for a language and, crucially, the identification of gaps and shortcomings that hamper the achievement of full DLE. This approach enables direct comparisons across languages, tracking their advancement towards the goal of DLE, and facilitates the prioritization of needs, especially to fill existing gaps. The metric is computed for each language on the basis of various factors, grouped into technological factors (technological support, e.g. available language resources, tools and technologies) and contextual factors (e.g. societal, economic, educational, industrial).

Our systematic collection of language resources, i.e. data (corpora, lexical resources, models) and LT tools/services for Europe's languages has resulted in more than 6,000 metadata records, which will be imported into the ELG catalogue and complement the existing, constantly growing inventory of ELG resources, thus providing information on the availability of more than 11,000 language resources and tools. All languages investigated by ELE are covered.

Using this collection as a firm empirical foundation for further investigation, we computed a DLE score for each language. We will present these results in full at the conference, but unsurprisingly, English was clearly shown as having the best context for the development of LTs and language resources. English is followed by German and French, and then by Italian and Spanish. After these five leading languages, variations between the configurations begin to be seen. Mostly, Swedish, Dutch, Danish, Polish, Croatian, Hungarian, Greek and Finnish are ranked in the upper half of the official EU languages. The official EU languages with the lowest scores are mostly Latvian, Lithuanian, Bulgarian, Romanian and Maltese.

Among the group of official national languages which are not recognised as official EU languages, Serbian is always the top performer, achieving a similar score to those of the lower-scoring official EU languages, while Manx is always presented as a downward outlier. Norwegian, Luxembourgish, Faroese and Icelandic achieve better scores than Albania, Turkish, Macedonian and Bosnian. The regional and minority languages are usually led by Saami South and Skolt.

These and other perhaps unexpected results will be explained at the conference. The results from our various surveys will also be shown, including the novel survey which targeted European citizens *per se*, where we look like surpassing 25,000 respondents from all over the continent.

3. Future Plans

ELE is on track to achieve its ambitious objectives with the consortium currently working on the SRIA which will be ready at the end of the project in June. The DLE metric has proven to be an extremely useful tool to demonstrate how prepared European languages are for the digital age, and what needs to be done to get them to the point where all such languages are digitally equal by 2030. As an extension of this work, we will soon publish our interactive DLE dashboard that makes use of the metadata records available in the ELG platform.

Acknowledgements

ELE is co-financed by the European Union under the grant agreement № LC-01641480 – 101018166 (ELE).

Reference

Georg Rehm, Federico Gaspari, German Rigau, Maria Giagkou, Stelios Piperidis, Natalia Resende, Jan Hajic, Andy Way. 2022. The European Language Equality Project: Enabling Digital Language Equality for all European Languages by 2030. *EFNIL Annual Publication Series*, Cavtat, Croatia (in press).