EUROPEAN LANGUAGE EQUALITY 2

D3.3

Extended database and ELE dashboard

Authors	Athanasia Kolovou, Maria Giagkou, Stelios Piperidis, Miltos Deligiannis, Leon Voukoutis (ILSP)
Dissemination level	Public
Date	30-06-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
Deliverable number	D3.3
Deliverable title	Extended database and ELE dashboard
Type	Report
Number of pages	13
Status and version	Final
Dissemination level	Public
Date of delivery	30-06-2023
Work package	WP3: Strategic Research, Innovation & Deployment Agenda: Maintenance and Extension
Task	Task 3.3 Extension and maintenance of the ELE dashboard
Authors	Athanasia Kolovou, Maria Giagkou, Stelios Piperidis, Miltos Deligiannis, Leon Voukoutis (ILSP)
Reviewers	Federico Gaspari (DCU), Marie Mattson (EFNIL)
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
CF	Contextual Factor
DLE	Digital Language Equality
ELE	European Language Equality
ELE1	European Language Equality (preceding project)
ELE2	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
EU	European Union
LR	Language Resource
LT	Language Technology
R&D	Research and Development
TF	Technological Factor

Abstract

In the frame of the follow-up 12-month European Language Equality (ELE2) project the European Language Grid (ELG) Catalogue was further expanded with approximately 3,000 new metadata records for datasets, software and projects. In parallel, the Digital Language Equality (DLE) dashboard was not only extended to reflect the expanded ELG catalogue collection of LRs, but it was also enhanced with new functionalities and data visualisation capabilities. These include the visualisation of evolution over time and of correlations between various factors for each language. These enhancements unveil supplementary data attributes that can be leveraged for monitoring the advancement of technology support across Europe's languages. They have the potential to spark inquiries and to stimulate further investigations into the level of technology support available for Europe's languages, and especially to provide evidence to initiate and guide much-needed interventions to support multilingualism in Europe enabled by Language Technologies (LTs).

1 Introduction

As part of the first ELE project (ELE1), a Digital Language Equality (DLE) monitoring and visualisation dashboard was developed (Giagkou et al., 2022). The dashboard is an outlet of the DLE metric and it provides digital readiness computations and comparative visualisations across languages. It uses as input the metadata descriptions of the European Language Grid (ELG) Catalogue resources (corpora, models, tools/services, etc.) and an appropriately linked database of the contextual factors that have been used to compute the DLE metric. Having a critical role as a monitoring mechanism and as an intuitive user-friendly presentation of the DLE metric scores, thus facilitating communication of the key ELE messages to stakeholders, the ELE dashboard was maintained and further extended in the frame of ELE2. This task entailed the following activities:

- Continued population of the ELG Catalogue with newly identified resources, with a combination of automated means (e.g., harvesting of external sources) and manual curation, including metadata mappings and conversions.
- Implementation of enhanced functionalities and visualisations of the dashboard, offering continuous monitoring of the DLE metric and the underlying up-to-date data.

The outcomes of the activities regarding the population of the ELG Catalogue are presented in Section 2, while Section 3 provides a description of the enhancements applied on the dashboard backend and user interface.

2 Database with newly identified language resources

The population of the ELG database and Catalogue has been a continuous activity of the ELE2 project. New metadata records have been created for newly identified language resources and technologies (LRTs, data and tools/services) as well as for Research and Development (R&D) projects in the area of Language Technology (LT). The identification of new LRs was achieved via either automated means, i.e. by harvesting relevant external catalogues and repositories (see Section 2.1 for more details), or through the ELE2 consultations, e.g. the feedback collected in particular through the online SRIA endorsement form and the meetings with various stakeholders. The new metadata records were thus imported to ELG either automatically, following appropriate mappings, conversions and deduplication, or, complementarily, via the manual ELG operations for LR contributions.

In total **2,832 new metadata records** were added to the ELG Catalogue **from July 1st, 2022** (the start date of the ELE2 project) **until May 31st, 2023** (a month before the end of the project). All the new metadata records created by ELE2 have been added to the ELG database and published on the ELG Catalogue, resulting – at the time of writing – in more than 17,000 records overall.[1]

The resources that have been added are further described in the following sections.

2.1 Data and software

During the lifespan of ELE2 2,319 LRs, including data and tools/services, have been added. Their distribution per resource type is presented in Table 1.

LR type	# new metadata records
Corpora	1,413
Lexical/conceptual resources	205
Language descriptions (including models)	60
Tools/services	641
Total	**2,319**

Table 1: Number of new LRs by resource type

A large number of the new metadata records (902) have been harvested, converted and mapped from Zenodo,[2] which was one of the external repositories targeted during this period (see Figure 1). Harvesting of ELRC-SHARE also continued and resulted in a significant number of new bi- / multi-lingual datasets (725). Several other sources, e. g. CLARIN repositories and the ELRA catalogue, also continued to act as seed repositories. On top of that, 444 metadata records have been created manually through the ELG metadata editor.

The new metadata records cover many different languages. Figure 2 shows the number of times each of the official EU languages appears as resource/input language. The pattern of prevalence of English, followed by German and Spanish, is evident again, as in our previous investigations on the datasets hosted in ELG until the end of 2021. Noteworthy on this graph is the comparatively significant number of new resources for a few low-resourced languages, such as Irish and Slovenian. Nonetheless, whether this leap in recent years is due to coordinated efforts for developing LRs for these particular languages or it is circumstantial, cannot be supported without additional evidence and in-depth explorations into the specific situations of the languages in question.

2.2 Projects

As a result of the ELE stakeholders consultation activities, we identified and documented 513 R&D projects in the field of LT and Language-centric Artificial Intelligence (AI). All of them are listed in the *Projects* section of the ELG Catalogue[3]. Many of these projects have delivered LRTs which are also catalogued in ELG. In such cases the appropriate relations were indicated in the metadata and the list of related resources is presented in the project description.

The majority of the projects have been supported by national and EU funds, as presented in Table 2.

[1] The ELG Catalogue can be browsed at https://live.european-language-grid.eu/catalogue/.
[2] https://zenodo.org
[3] https://live.european-language-grid.eu/catalogue/?entity_type__term=Project

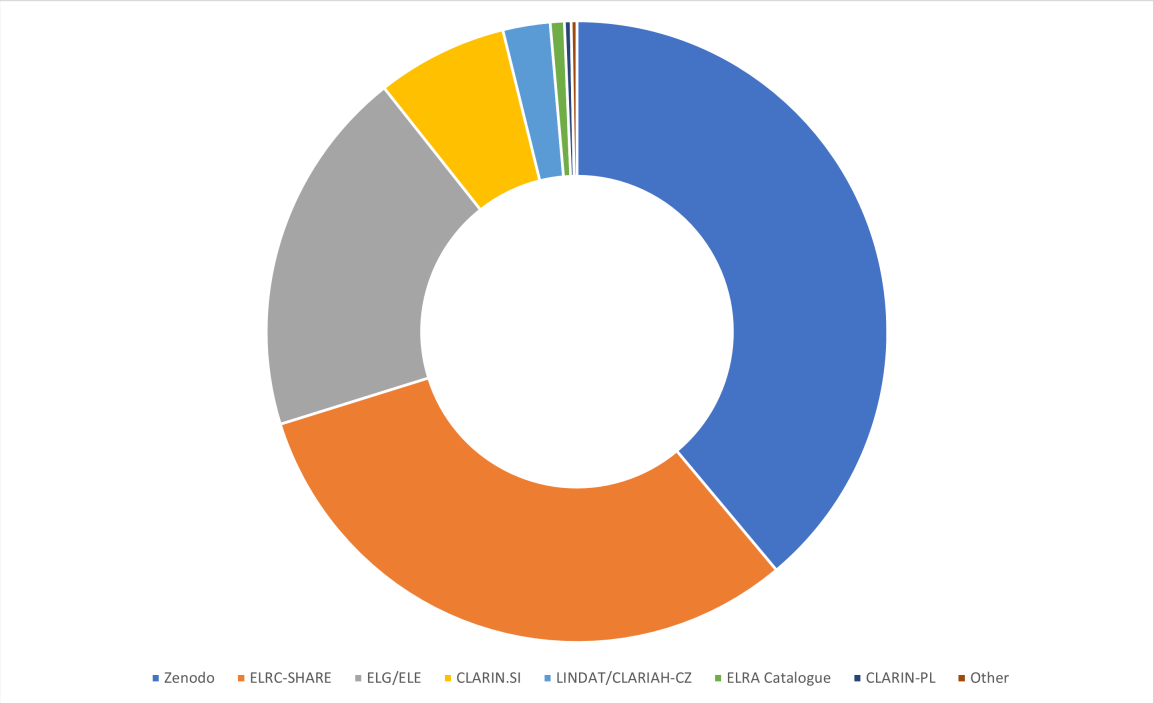

Figure 1: Source repositories of the new metadata records

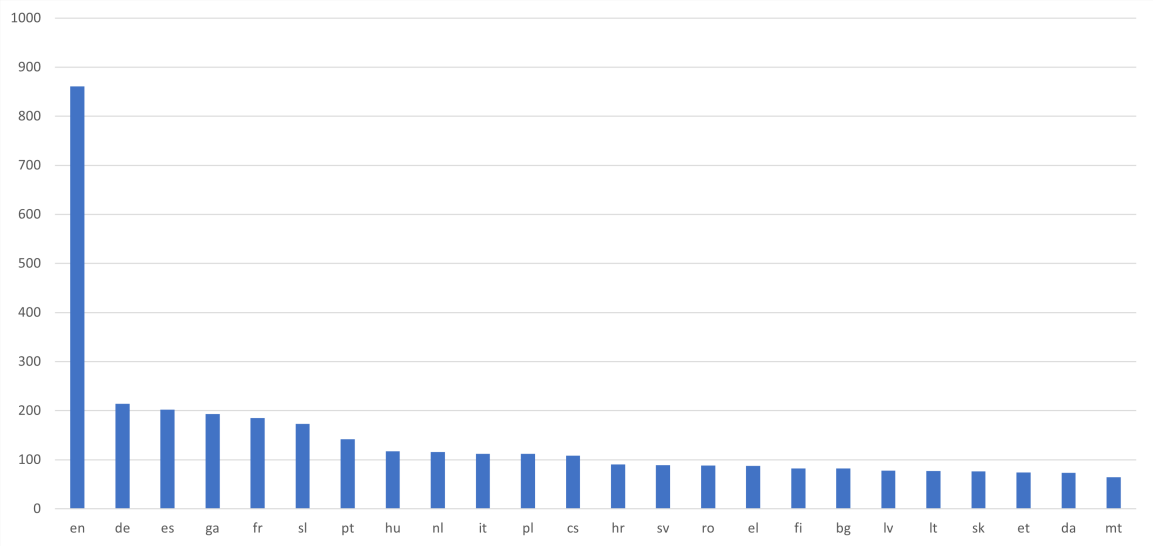

Figure 2: EU official languages' presence in the new metadata records

Identification and documentation of LT-related projects will continue beyond the ELE project lifetime, as the ELG Catalogue will continue to organically grow over time in a sustainable manner.

Funding type	# projects
National funds	329
EU funds	118
Regional funds	33
Unspecified	21
Other	19
Own funds	12

Table 2: Type of funding

3 Maintenance and extension of the ELE Dashboard

In order to provide a mechanism for dynamically exposing and monitoring the support that European languages receive through technology, ELE has designed and implemented an interactive dashboard that is available as part of the ELG platform.[4]

The dashboard is based on the ELG database and provides an overview of the DLE metric (Gaspari et al., 2022), and the two main components contributing to it, i. e., the technological (TFs) and contextual factors (CFs). The dashboard exposes the TFs (based on the contents of the ELG catalogue) and the CFs as interactive visuals dynamically created in response to user queries. With regard to the TFs, as the ELG catalogue organically grows over time, the dashboard provides an up-to-date overview of LT support that each language enjoys, also showing where the status is less than ideal or not at the expected level.

The first version of the user interface of the ELE dashboard consisted of three entry points (sections) (see Giagkou et al. (2022) for details). The first section displays the bar charts of the DLE metric for CFs and TFs for the languages selected by the user. The other two sections enable users to dive into a more detailed comparison of a subset of the TFs, both across languages and within a language. Datasets can be compared to software resources and, when selecting one of the two, a number of features of the corresponding resource class can be compared, e. g. monolingual vs. bilingual corpora.

Despite the fact that the initial version of the dashboard offered valuable insights into the technological support for Europe's languages, additional evidence-based measures or indicators, in conjunction with the DLE metric, have been implemented. In the updated version of the dashboard, developed as part of ELE2, the *Cross-language comparison* part of the dashboard has been enhanced with two new subsections, each providing new table- or chart-based comparative visualisations of languages:

- Heatmaps and tables
- Radial bar

From a user experience perspective, these additions are presented as options under the main menu item *Cross-language comparison*, complementing the already available option for comparisons per resource types (corpora, tools, lexical resources, etc) and resource features (e. g. media types, access rights, etc.) with histograms (bar charts) (see Figure 3).[5]

The *Heatmaps and tables* section exposes the ELG database source data in the form of either actual number of resources or percentages representing the contribution of each language per resource type/function, i. e.

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

[5] The *Histograms* section is not presented in this document, as it was implemented in the frame of the ELE1 project and is detailed in (Giagkou et al., 2022).

- for datasets: Corpus, Model, Lexical/conceptual resource, Grammar, Uncategorised language description;
- for software: Text Processing, Speech Processing, Information Extraction and Information Retrieval, Image/Video Processing, Translation Technologies, Natural Language Generation, Support operation, Human Computer Interaction and Other functions of software.

These measurements are visualised in tabular form or in the form of heatmaps. The intensity of colors on the heatmap indicates in an intuitive but consistent manner the magnitude or density of the data being represented (Figure 4). The generated table can be downloaded by the user as a CSV file, while the heatmap visual can be downloaded locally as a SVG file.

In the *Radial bar* section users can generate data visualisations of a circular layout, with bars radiating out from the centre of the circle. Each bar represents a language and the length of the bar corresponds to the number of available resources (all datasets per type and all software per function) in the ELG catalogue (Figure 5). The radial bar chart allows for simultaneous comparison of multiple languages, making it especially helpful for illustrating the contribution of various resource types to the total per selected language. The generated radial bar can be downloaded as a svg file.

One of the most signifant enhancements to the dashboard in the frame of ELE2 is the consideration of the perspective of time and the possibility to compare languages as per the evolution of the number of resources over time and to monitor their evolution thereof.

The *Evolution over time* section enables users to create charts displaying either a) the overall data evolution over time, or the b) the intensity at which data evolved over each time period. The overall evolution of the number of resources displays data as a series of points

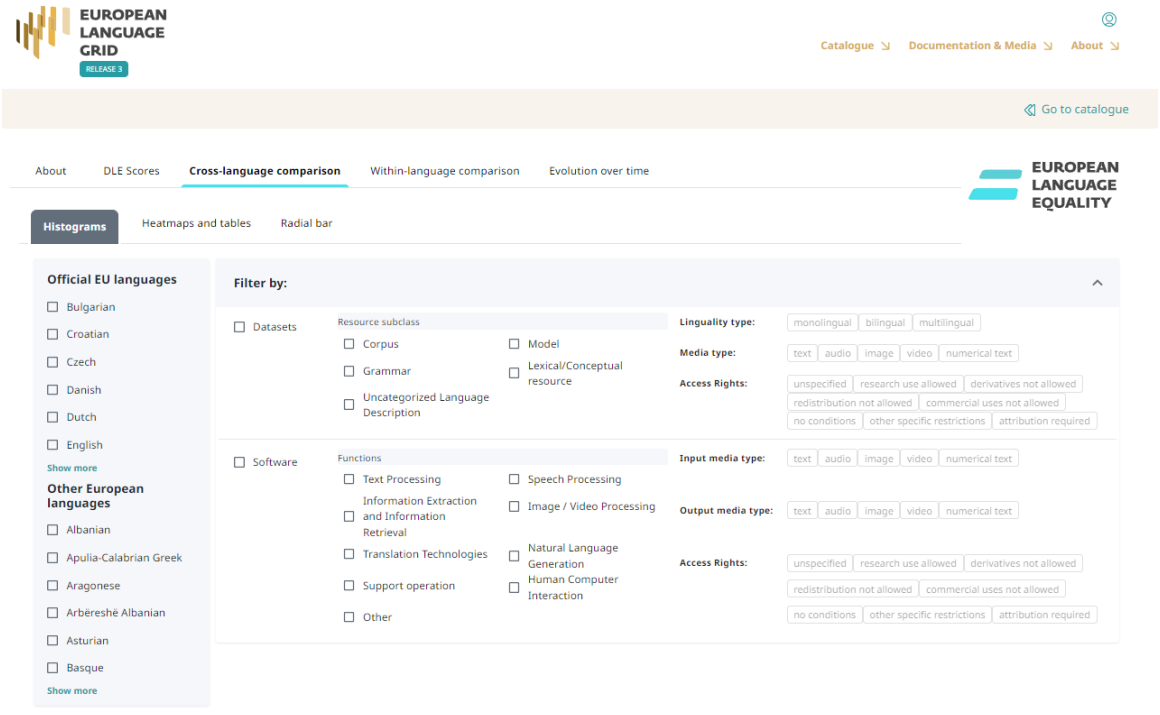

Figure 3: Entry points for cross-language comparison per resource types and features: histograms, heatmaps and tables and radial bars

connected by a line (Figure 6). The data is plotted on a coordinate system, with the horizontal axis representing time and the vertical axis representing the number of resources. Multiple languages can be selected for simultaneous comparison and the time periods can also be viewed per semester or per quarter, depending on the granularity of the visualisation and analysis that is required by the user. The charts also include additional features such as labels, axes, grid-lines, and legends to enhance readability and facilitate data interpretation in a user-friendly manner.

The *Intensity evolution* chart provides an overview of the rate at which data for each language has been created or imported to the ELG Catalogue over time. Again, multiple languages can be simultaneously selected for comparison and the time periods can also be viewed per semester or per quarter (Figure 7). In this chart, the data for each language is plotted on a vertical axis, with time on the horizontal axis. Each language is represented by a separate line that is drawn as a filled-in area. The areas are arranged in such a way that the language that has the highest number of resources appears wider. As time progresses, the areas become wider or thinner, reflecting changes in the relative positions of the language's contribution. This type of chart (also called "area bump chart"), can be particularly useful when comparing multiple languages selected by the user, tracking relative changes and positioning over time, e.g. as a result of dedicated funding and policy interventions (or, rather, in case of lack thereof).

These additions include evolution over time and correlations between various factors for each language. The updated visuals unveil supplementary data attributes that can be leveraged for monitoring the advancement of technology support across different languages. They have the potential to spark inquiries, stimulate further research, and offer insightful ideas to individuals, domain experts, as well as policy- and decision-makers, at the European, national, regional and local levels.

Architecturally, the ELE dashboard consists of two layers. The ELG database provides the source data to be exposed, in particular the source data for the TFs. The ELG database con-

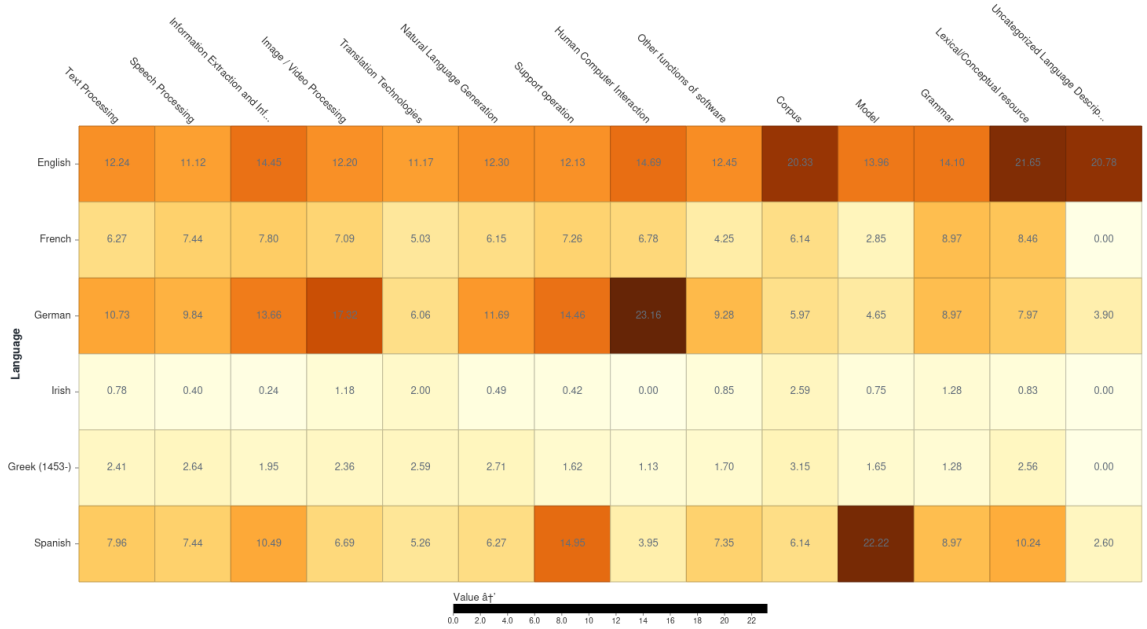

Figure 4: Example of a heatmap: contribution of English, French, German, Irish, Greek and Spanish per resource type (percentages)

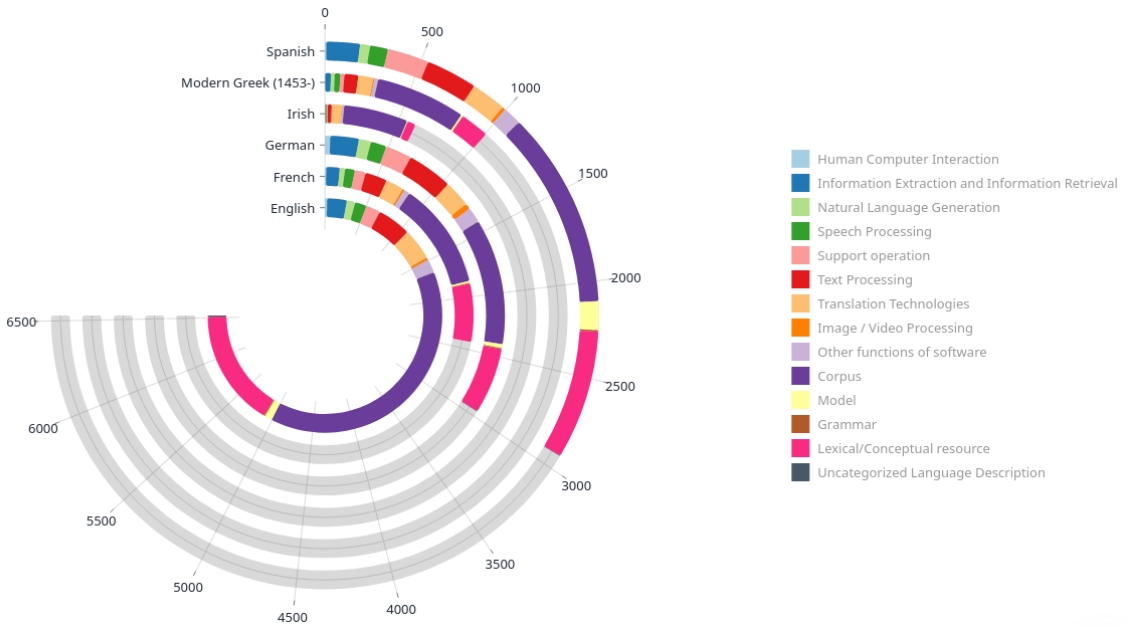

Figure 5: Example of a radial bar comparing English, French, German, Irish, Greek and Spanish

tents are indexed and saved in appropriate JSON structures. Each user query retrieves the respective results from JSON and exposes them to the front-end. All results are visualised as interactive graphs. For the front-end implementation, multiple libraries are utilized, i.e. the react-chartjs-2 library[6] for charts, Nivo library[7] for the heatmap, radial bar and time evolution charts, while the table view is created using the Data Grid component from Material UI.[8] In addition, the chartjs-plugin-zoom library[9] is used for additional features like pan and zoom options on a selected chart. Finally, after creating a chart with their preferred options on the dashboard, users have the option to download them in vector-based SVG format.

4 Conclusions

The DLE dashboard is one of the project's main outputs and a key contribution to its goals. The ability to visualise the LT support that each European language enjoys in real time provides an effective and comparative means to demonstrate the LT level of groups of languages and resources. The DLE dashboard would not be implementable if not fed with the data hosted in the ELG Catalogue. At the time of writing (June 2023), after its continuous population with newly identified LRs by ELE, the ELG catalogue comprises more than 17,000 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. Considering the current dynamicity, we consider the present status of the ELG repository adequately representative with regard to the current existence of LT resources for Europe's languages.

The recent updates to the dashboard, in the form of heatmaps, radial bars, and evolution

6 https://react-chartjs-2.js.org/
7 https://nivo.rocks/
8 https://v4.mui.com/components/data-grid/
9 https://www.chartjs.org/chartjs-plugin-zoom/latest/

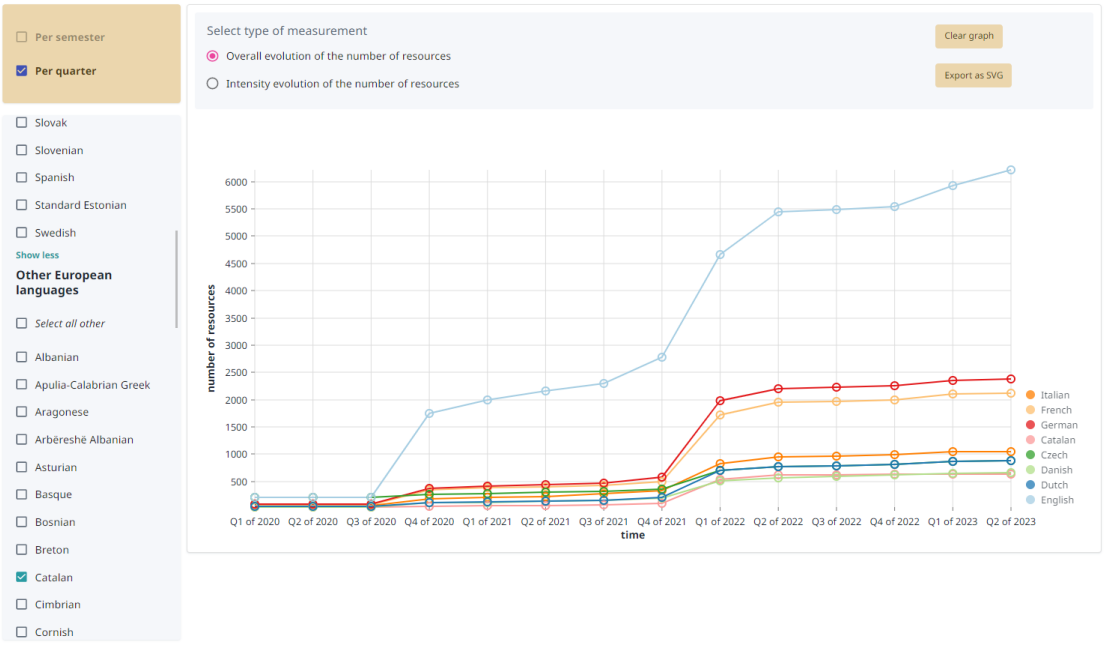

Figure 6: Evolution over time

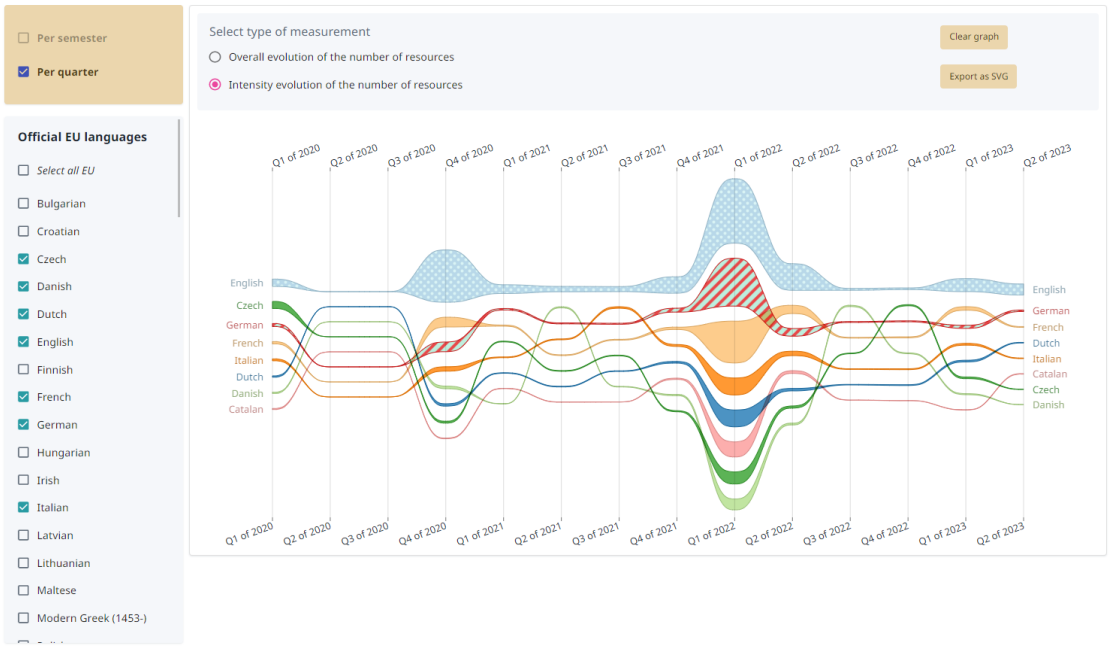

Figure 7: Intensity evolution

over time, add new possibilities for cross-language comparisons which can be utilised to uncover the main types of LRs that are currently lacking in order to expose the gaps that must be filled in the near future in the interest of achieving full DLE in Europe (see ELE2 deliverable D3.2).

The updated visuals unveil supplementary data attributes that can be leveraged for monitoring the advancement of technology support. They have the potential to spark inquiries, stimulate further research, and offer insightful ideas to individuals, domain experts as well as decision- and policy-makers all across Europe.

References

Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way. Deliverable D1.3 Digital Language Equality (full specification), 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_3.pdf. 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 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.