EUROPEAN LANGUAGE EQUALITY

D1.37

Database and Dashboard

Authors	Maria Giagkou, Penny Labropoulou, Stelios Piperidis, Miltos Deligiannis, Athanasia Kolovou, Leon Voukoutis
Dissemination level	Public
Date	30-04-2022

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	D1.37
Deliverable title	Database and Dashboard
Type	Report
Number of pages	22
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 30-04-2022 – Actual: 30-04-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Maria Giagkou, Penny Labropoulou, Stelios Piperidis, Miltos Deligiannis, Athanasia Kolovou, Leon Voukoutis
Reviewers	Federico Gaspari, Katrin Marheinecke, Annika Grützner-Zahn
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 Figures

List of Tables

List of Acronyms

DLE	Digital Language Equality
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
LR	Language Resource/Resources
LRTs	Language Resources and Technologies
LT	Language Technology/Technologies
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NLU	Natural Language Understanding
SRIA	Strategic Research and Innovation Agenda
TRL	Technological Readiness Level

Abstract

Through an unprecedented large-scale collaborative activity, more than 40 ELE research partners, who are experts in more than 30 European languages, collected and aggregated metadata for more than 6,000 Language Resources and Technologies (LRTs). The results of this data collection have been integrated into the European Language Grid (ELG) further enriching its own catalogue and providing a detailed, empirical and dynamic map of technology support for our languages. The level of technological readiness and the various factors that contribute to Digital Language Equality (DLE) can be discovered, browsed and further investigated by means of comparative visualisations across languages through a web-based dashboard application. The ELE dashboard supports querying and visualising the empirical data that we consider components of DLE as well as monitoring and comparing the evolution of technological support within and across languages, advancing towards the goal of digital equality for all European languages by 2030.

1 Introduction

This deliverable reports on part of the results of ELE Task 1.3, which sought to investigate the state of play of Europe's languages in 2021/2022 with regard to their digital readiness. The task was based on a comprehensive metadata collection activity. Originally informed by the working definition and indicators of language equality (Gaspari et al., 2021, 2022) and by the state-of-the-art report (Agerri et al., 2021), T1.3 discovered and appropriately documented data sets and tools/services pertinent to Language Technology (LT) for each language.

The findings of this investigation per language have been reported in the ELE language reports series (D1.4 to D1.36). The report at hand (Deliverable D1.37) presents the results of the ELE metadata collection activity and the software developed to facilitate digital readiness computations and comparative visualisations across languages, highlighting strong and weak points of the digital support offered to each of the languages addressed by ELE.[1]

Section 2 reports on the collected LRTs that resulted from our the collaborative expert-based, large-scale metadata aggregation activities. It also presents the basic curation and processing tasks that were applied to convert their metadata into an ELG-compatible metadata schema and then import them into the ELG database. Section 3 presents the web-based ELE dashboard, developed as part of the ELG platform. The ELE dashboard exposes and visualises the descriptive statistics of the LRTs hosted on the ELG database and catalogue as well as the technological and contextual scores of the DLE metric, as defined in Gaspari et al. (2022). Finally, by way of conclusion, Section 4 offers an outlook on the future uses and relevance of the dashboard going forward, including beyond the lifetime of the ELE project.

2 Database of Empirical Data

In close collaboration with its sister project, the European Language Grid (ELG), ELE made use of the ELG Platform and Catalogue, which offered the empirical base for the technological factors of the Digital Language Equality metric. This decision is based on the fact that the ELG Catalogue is Europe's most comprehensive list of LRTs. On top of its own LRTs, ELG harvests several major LR/LT repositories.[2] To make sure that the ELG Catalogue is adequately representative of the status of technological readiness of European languages, the ELE project completed a comprehensive and large-scale review study of the level of support

[1] https://european-language-equality.eu/languages/
[2] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

the European languages receive through Language Technology (LT). In particular, at least 40 different organisations, either ELE consortium partners or collaborating through partners' networks, acted as language informants for their language and collectively they investigated, discovered and appropriately documented components that contribute to the level of technological support of European languages. These components are resources that can be used for the development of LT, LT tools and services, research groups and industry providers of LT, projects in the framework of which the resources and tools have been developed, etc.

The ELE metadata collection process focused on LRTs for the official, co-official, regional, minority and community European languages for which the ELE consortium and its collaborating network had expertise. These LRTs included:

- corpora or datasets: collections of raw or annotated, monolingual or bi-/multilingual, mono- or multimodal, text segments or documents, audio transcripts, scripts, audio and video recordings, etc., as well as learner corpora and sign language corpora.
- language descriptions, comprising language models and computational grammars.
- lexical/conceptual resources, comprising computational lexica, terminological databases, gazetteers, ontologies, term lists, thesauri, etc.
- tools and services: services offered through the web, other networks or running in the cloud, but also downloadable tools, source code, etc. These include basic NLP tools for the European languages (morphological analysers, POS taggers, lemmatizers, parsers, etc.), authoring tools (e. g. spelling, grammar and style checkers), tools/services for information retrieval/extraction/mining, text and speech analytics, machine translation, natural language understanding and generation, speech technologies, conversational systems, etc.

The ELE informants were asked to document only resources that were not already included in the ELG Catalogue. Three options were provided to ELE informants as instruments for the documentation of resources they discovered: (i) the ELG metadata editor (Figure 11 in Appendix), and/or (ii) an online form (Figure 12 in Appendix), and/or a spreadsheet which was automatically populated by the responses to the online form and editable by the informant for direct manual recording of new resources. The last two items, the online form and the spreadsheet, were appropriately configured to render a very simplified version of the ELG metadata schema (Labropoulou et al., 2020).

By adhering to and utilising the ELG metadata schema, interoperability with the ELG Catalogue of LRTs was guaranteed, thus allowing for the aggregation and ingestion of the LRTs documented by the ELE partners into the ELG Catalogue, in an as automated as possible manner. On the other hand, having set as a priority the documentation of as many LRTs as possible over a detailed documentation for each of them, and in order to respond to the variety of sources from which the ELE informants could discover relevant information, only a subset of the ELG metadata categories have been included in the ELE online form. These were carefully selected to elicit sufficient information for the purposes of ELE.

The online form contained the following metadata categories (elements marked with an asterisk were mandatory):

- identification: resource type*, resource name*, resource short name, landing page*, description*, publication year, resource provider (organisation name)
- contact data: name & homepage of source, contact email
- classification: keyword, domain
- funding information: funding project & funding type

- usage information: licence, access rights
- technical information for data resources: subclass*, language* and, where applicable, geographical variety, multilinguality type, media type*, size; in addition, for annotated corpora, annotation type, and, for lexical/conceptual resources, encoding level
- technical information for tools/services: function*, Technological Readiness Level (TRL), whether they are language independent*, and if not, the language and, where applicable, geographical variety of the input resource, media type of the input resource, and language, geographical variety and media type of the output resource.

Recommended controlled vocabularies, in the form of lists of values from which users could select a value, were used where possible (e. g. for language), yet informants could also add free text values. Depending on the element, adding multiple values was possible (e. g. for domains, languages, keywords, etc.). Mandatory elements were marked as such with validation imposed.

The records created by the ELE informants were curated before being imported to the ELG database. The curation process involved manual and (semi-)automatic processing of the metadata and indicatively included:

- Surface deduplication at the level of resource name, resource short name and landing page. Some records were identified as duplicates of existing ELG records while some were duplicates of other ELE informants' contributions. Sometimes duplicate records contained different or contradicting values, e. g. different functions, licences, etc. In such cases the source had to be consulted and the record manually corrected.
- Completion of mandatory metadata, where possible. The ELG metadata schema includes some mandatory elements, e. g. resource name and media type. Missing values in mandatory metadata result in invalid records and failure to import to the database. To minimise loss of data due to missing mandatory values, we consulted the source and filled in missing metadata where possible. For remaining records, when the data type of the element permitted this, we used the value "unspecified".
- Harmonisation and mapping of metadata values. In many cases, metadata values were provided by the informants as free text. Where possible and appropriate, these were semi-automatically mapped to controlled vocabularies. For instance, values such as "speech synthesis", "speech synthesizer", "text to speech", "TtS" were all mapped to "Speech synthesis". In certain cases the value provided by the informant was not mappable to an ELG controlled vocabulary (e. g. "200+ languages" for the language metadata element). In such cases, where possible, the source had to be consulted and the record was manually corrected.
- Grouping of values into meaningful equivalents or supersets. For instance, the domain values "journalism", "journalism and publishing" and "journalistic" were all mapped to the single value "journalism and publishing", while tool functions "speech synthesis" and "speech recognition" were additionally grouped under the functions group "speech processing", etc.

This systematic collection and careful curation of language resources for Europe's languages by the ELE partners resulted in 6,790 metadata records. Out of these, approximately 400 records were eventually discarded during metadata curation and processing (mainly because of duplicates or incomplete mandatory metadata), resulting in **6,362 records for import to the ELG database**. These included 2,215 metadata records describing LT tools and services as well as 4,147 records describing data resources, i. e., corpora, lexical/conceptual

resources and language descriptions, grammars or language models (see Figure 1 for details of the breakdown). They covered all the languages addressed by the ELE language reports series (i. e. ELE deliverables D1.4–D1.35), as presented in Figure 2.

	Number of records
Datasets	4,147
Tools and services	2,215
Total	6,362

Table 1: Number of LRTs identified by ELE language informants and imported into ELG

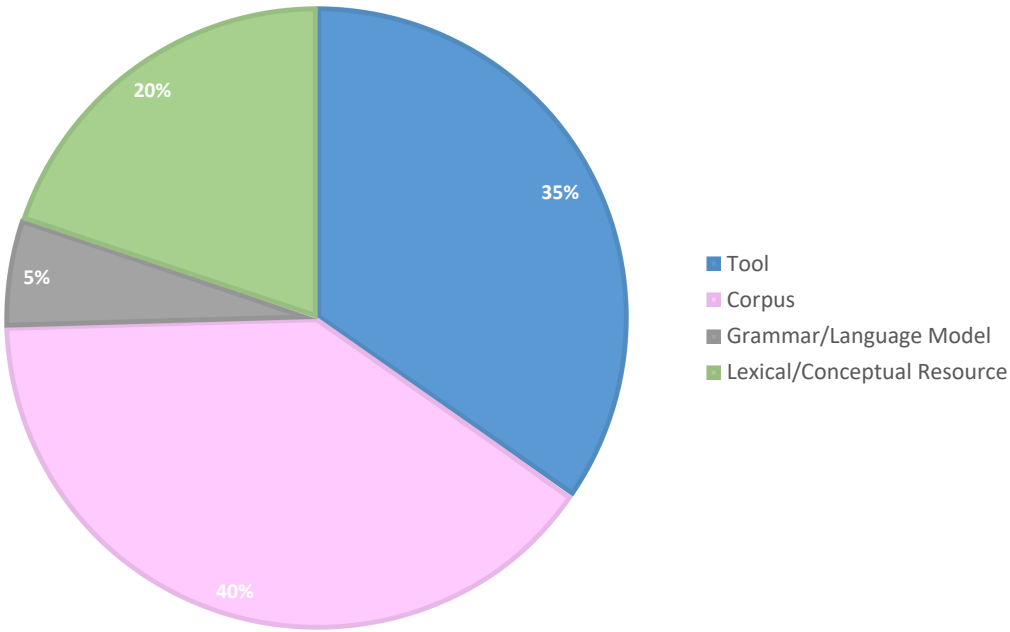

Figure 1: Distribution of LRTs identified by ELE language informants and imported into ELG

This set of metadata records was imported into the ELG Catalogue, complementing ELG's existing collection of LRTs. Currently (April 2022), after the addition of the records created by the ELE partners, the ELG Catalogue comprises 12,033 LRTs, as presented in Table 2.

This consolidated collection of LRTs of the ELG Catalogue is searchable and browsable through the ELG Catalogue interactive interface and it is exposed and visualised on the dashboard described in Section 3.

3 Dashboard

In order to provide a precise and easy-to-use mechanism for exposing and monitoring the technological and contextual factors that contribute to DLE (Gaspari et al., 2022), we designed and implemented a web-based interactive dashboard as part of the ELG Platform.[3]

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

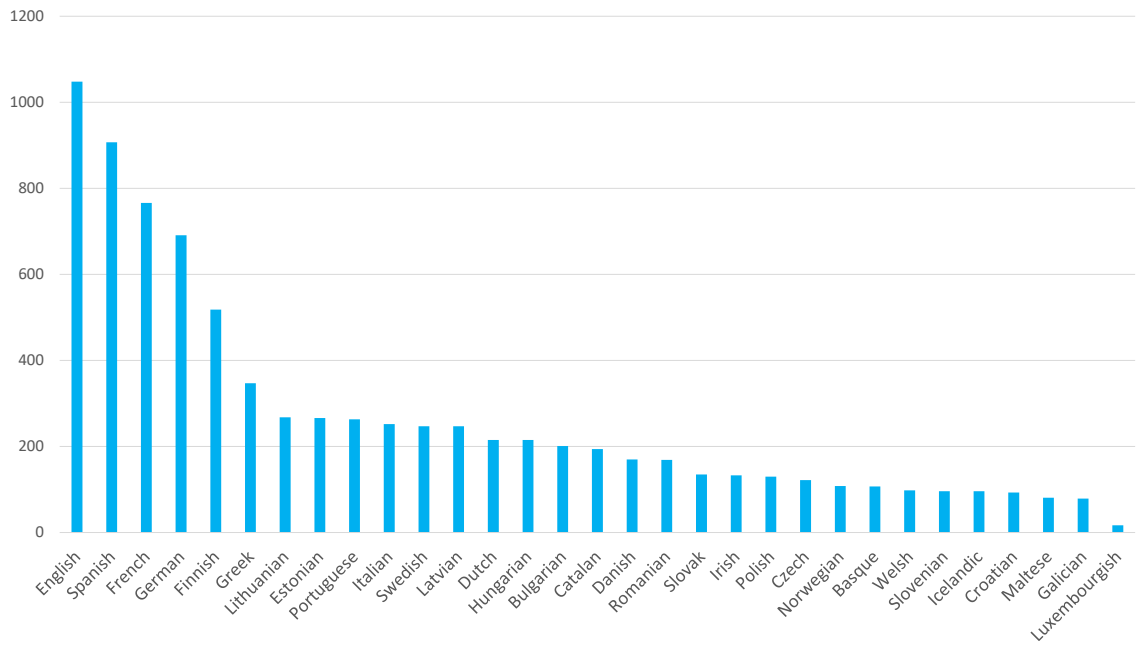

Figure 2: Number of LRTs identified by ELE language informants per language addressed by the ELE language reports series

	Number of records
Corpora	6,260
Lexical/conceptual resources	2,238
Language descriptions	433
Tools/services	3,102
Total	12,033

Table 2: Number of LRTs in the ELG database (April 2022)

The dashboard exposes the contents of the ELG database as interactive visuals dynamically created by user queries, thus providing a constantly up-to-date and consistent (i. e., comparable) measurement of the level of LT support and provision that each European language enjoys. The dashboard provides the figures, statistics and graphs, as appropriate, for:

- the technological and contextual scores of the DLE metric, calculated according to the detailed technical description presented in Gaspari et al. (2022);
- LRTs hosted on the ELG Catalogue, which constitute the source/base data for the technological factors that contribute to DLE.

3.1 Architecture and Implementation

Architecturally, the ELE dashboard consists of two layers: the database of the ELG Catalogue and the front end.

The ELG database contents are indexed and saved in appropriate JSON structures. Each user query retrieves the respective results from JSON and exposes them to the front end.

Specifically, the **scores of the technological part of the DLE metric** are calculated as follows:

1. Each LRT in the ELG database (dataset or tool) obtains a score ($Score_{LRT}$), which is equal to the sum of the weights of its relevant features.[4] Specifically for features Annotation Type and Domain, instead of simply adding the respective weight, the weight is multiplied by the number of unique feature values possessed by the LR in question.

   Example: Suppose a LRT in the ELG catalogue (LRT1) has the following features: corpus, annotated, monolingual, with three different annotation types (morphology, syntax, semantics), with text as media type, covering one domain (e. g. finance), with conditions of use research use allowed. Then, using the weights proposed in Gaspari et al. (2022), LRT1 is assigned the following score:

$$Score_{LRT1} = 5 + 1 + 2.5 + (3 * 0.25) + 1 + (1 * 0.3) + 3.5 = 14.05$$

2. To compute the technological DLE score for language X ($TechDLE_{LangX}$) one sums up the $Score_{LRT}$ of all LRTs that support language X (LRT1, LRT2, ... LRTN), i. e.

$$TechDLE_{LangX} = \sum_{i=1}^{N} Score_{LRTi}$$

While the technological factors are calculated dynamically and reflect the status of the ELG Catalogue's database at the time of accessing the dashboard, in the current implementation the **contextual factors** have been calculated offline, stored in a separate file and exposed to the respective tab of the dashboard front end.

With respect to the front end implementation, the challenge was to create a generic component that can manage and display the results of a number of possible user queries on the ELG database and that is simple enough to be maintained and reused over time. Moreover, a graphics library that can mount different types of graphics with the same data structure received from the back end was necessary. This condition was of utmost importance and was the determining factor when selecting the library for the task. Taking the above into

[4] The weights assigned to each feature and the associated values are presented in detail in Gaspari et al. (2022).

consideration, the react-chartjs-2[5] library for charts and the chartjs-plugin-zoom[6] library for features like pan and zoom functionalities on charts have been selected. For the creation of the interactive map we are using the *react-leaflet*[7] library, which is based on Leaflet.[8]

3.2 User Interface

The ELE dashboard web user interface can be accessed at

https://live.european-language-grid.eu/catalogue/dashboard.[9]

From a user experience perspective, there are four entry points (sections). The first section (*DLE scores*) displays the bar charts of the technological (Figure 3) and the contextual scores (Figure 4) of the DLE metric for the language(s) selected by the user. The user can select and compare the scores of any number of languages, up to a maximum of 92 of them, which is the total number of languages addressed by the ELE project. The languages list is provided in the form of a multiple choice list, grouped into *EU official* and *Other European languages*, alphabetically ordered within each group. The selected languages which are presented on the two bar charts, technological and contextual, are sorted in descending order of the respective DLE scores.

In the other two sections users can dive into a more detailed comparison of a subset of the technological factors across languages and within a language (*Cross-language comparison* and *Within-language comparison*, respectively). Through these two sections the ELG database can be queried and the respective statistics and graphs of the numbers of LRTs for specific languages in the ELG database can be generated. The number of datasets vs. software resources can be compared across languages, as in the instance shown in Figure 5.

Additionally, languages can be compared within each resource class with respect to different dimensions. For datasets, one or more resource subclasses (i. e. corpus, lexical/conceptual, grammar, model or other language description) can be compared, while for software, the user can generate graphs for the different function groups performed by the tool/service (i. e. Text Processing, Speech Processing, Translation Technologies, Image/Video Processing, Human Computer Interaction, Natural Language Generation, Information Extraction and Information Retrieval, Support operation and Other). The query can be further refined by applying a number of features, characteristic of the corresponding resource class. For datasets, these are the linguality type, the media type and the access rights. For software, the available query criteria are input and output media types and access rights. For instance, Figure 6 presents the graph generated out of a query for software for Text processing or Speech processing or Translation technologies, which is not provided for commercial use, for Bulgarian, Croatian, Czech, Danish and Dutch.

The *Within-language comparison* functionality allows for further investigations of the availability of LRTs for a single language. The functionalities are almost identical to those of the *Cross-language comparison* section with a few differences, e. g. the use of clustered bar charts where appropriate (for example Figure 7) and the breakdown of bars per resource subclass or function (e. g. Figure 8).

Finally, the *Map* section (Figure 9) visualises part of the situational context of European languages: The user can select a language addressed by ELE, and the countries where this language is spoken are highlighted on the map. To maintain focus on the European context

[5] https://react-chartjs-2.js.org
[6] https://www.chartjs.org/chartjs-plugin-zoom/
[7] https://react-leaflet.js.org
[8] https://leafletjs.com
[9] Should this URL become inactive in the future, the new link to the ELE dashboard will be provided in the ELG platform homepage.

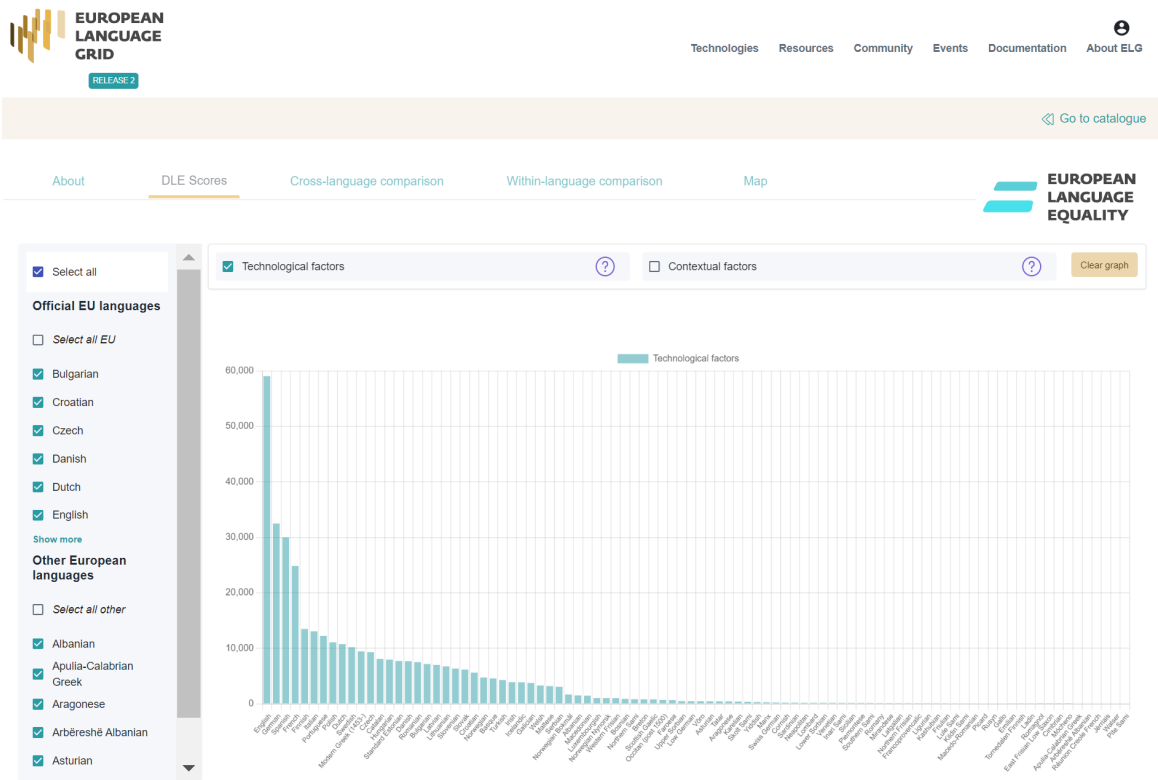

Figure 3: Dashboard: Technological DLE scores for all languages addressed by ELE (as of 25/04/2022)

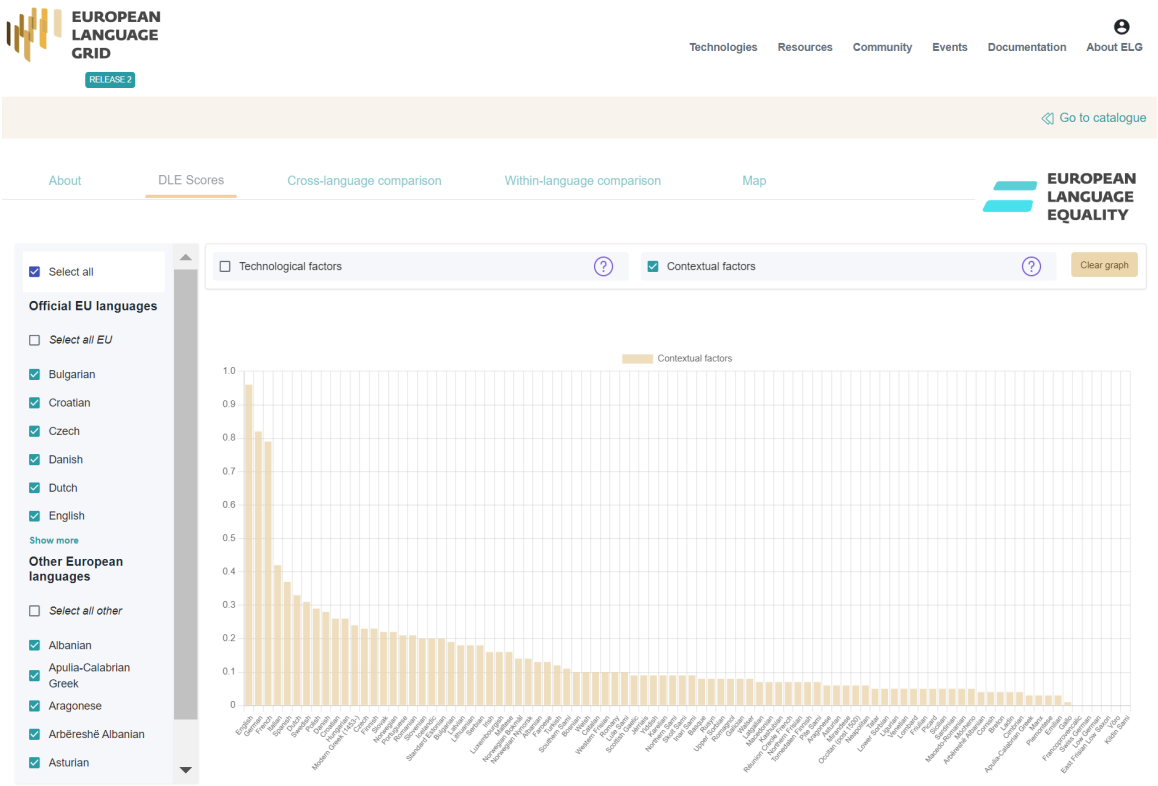

Figure 4: Dashboard: Contextual DLE scores for all languages addressed by ELE

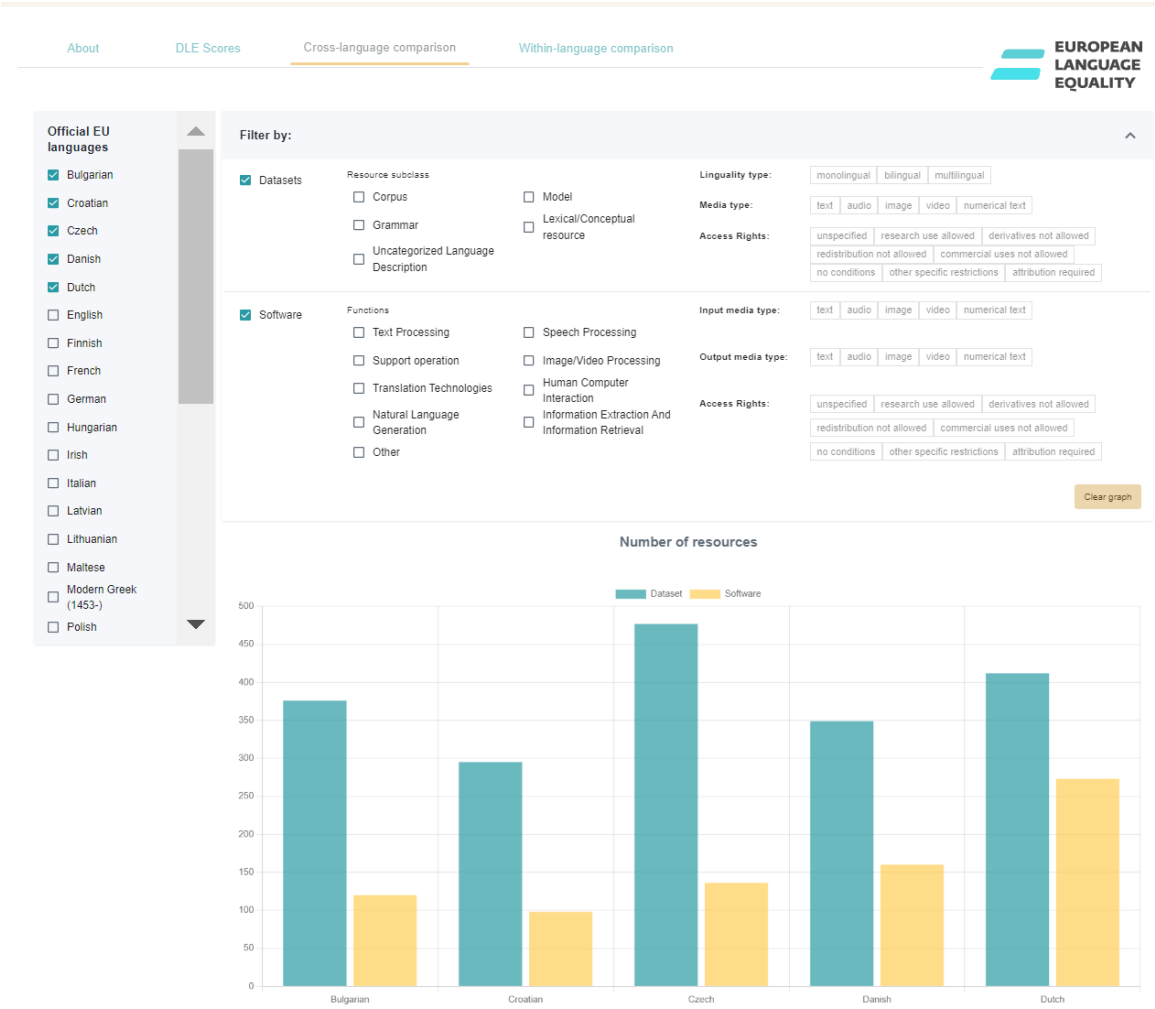

Figure 5: Dashboard: Comparing the numbers of datasets and software across five languages

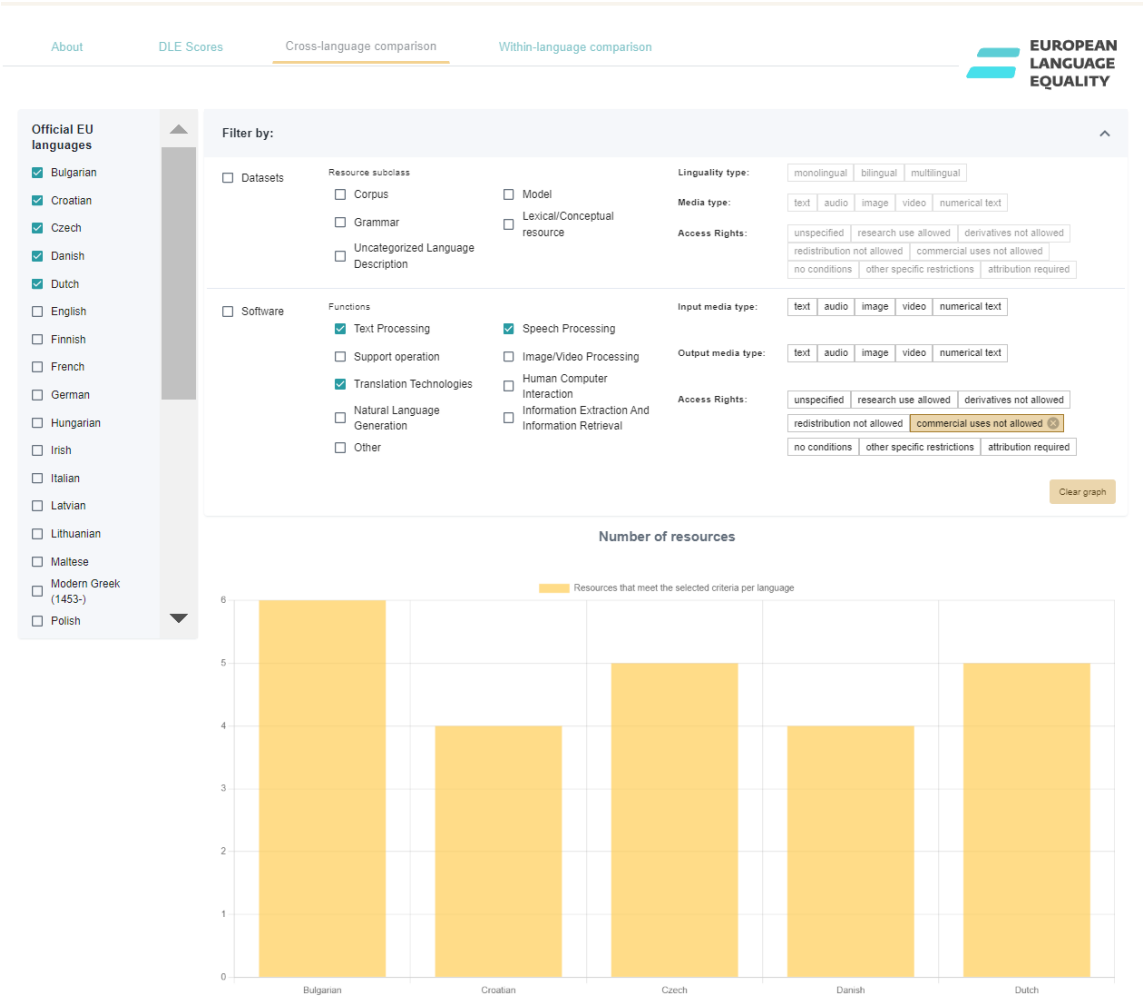

Figure 6: Dashboard: Example graph – software with specific functions and access rights

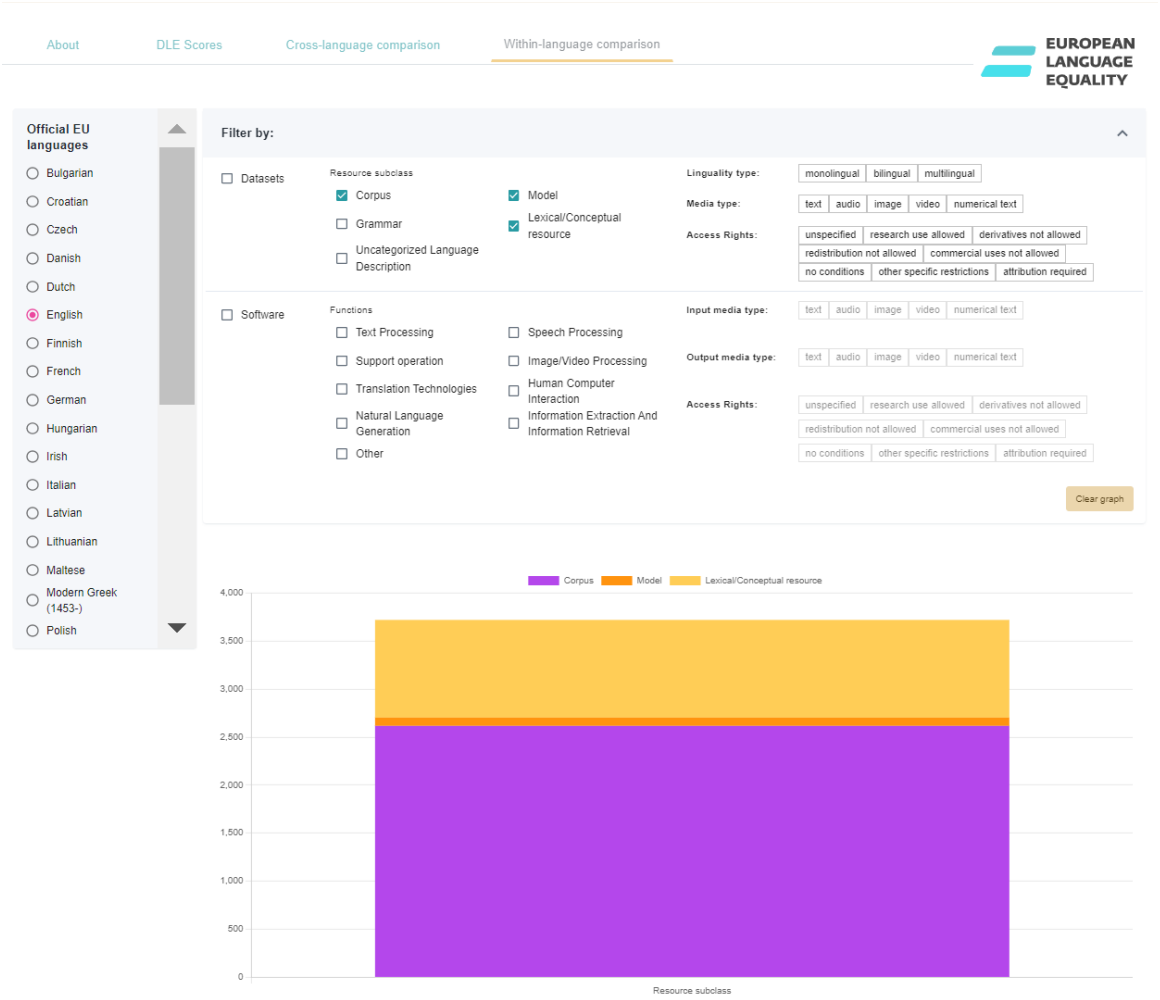

Figure 7: Dashboard: Example graph – comparing the number of corpora, models and lexical resources for English

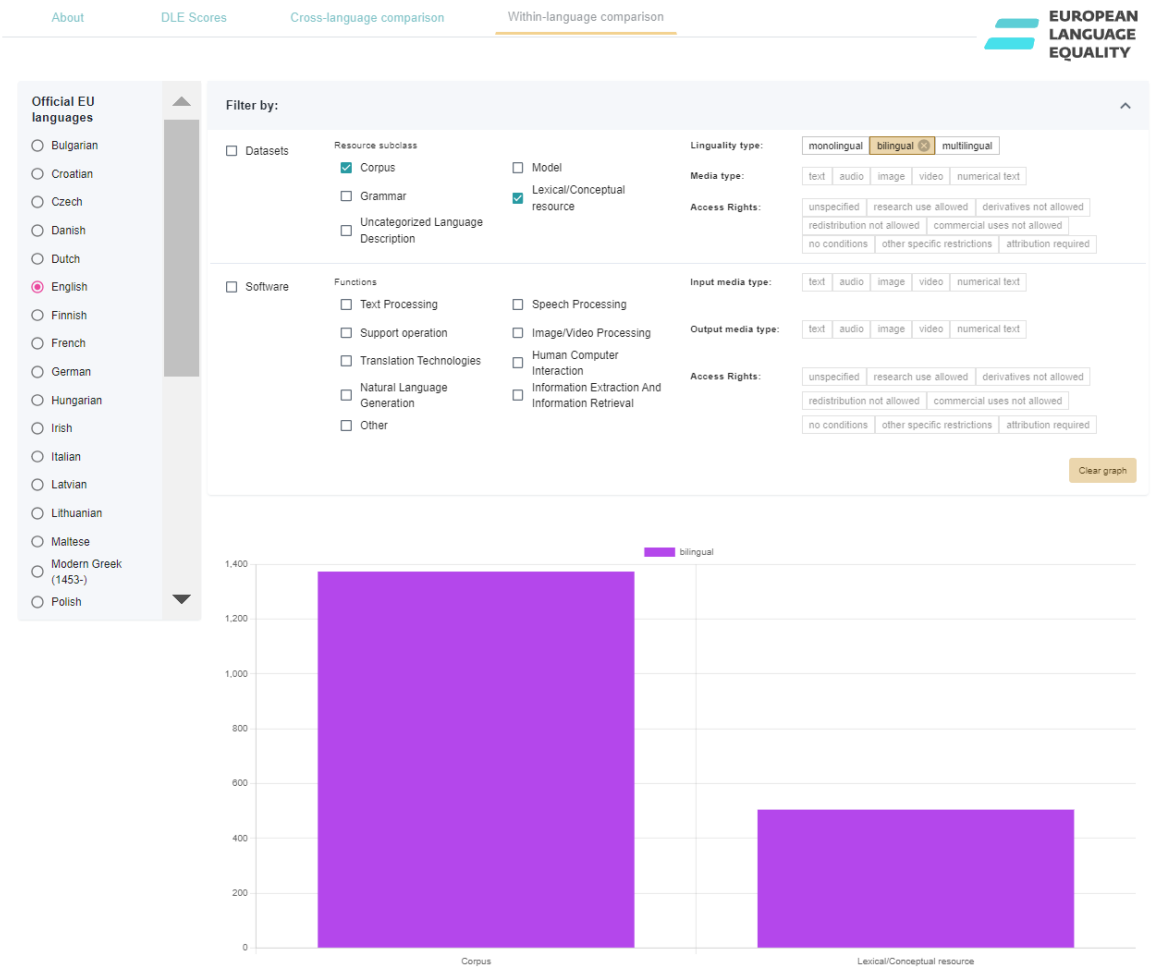

Figure 8: Dashboard: Example graph – comparing the number of bilingual corpora and bilingual lexical resources for English

and on the countries represented in the ELE consortium, only EU member states, EEA countries and Switzerland are active on the map. By clicking on a specific highlighted country, the numbers of L1 and L2 speakers of the language in the country are presented (Figure 10). The information about the number of speakers per country was mainly derived from Ethnologue[10] (September 2021), while for some languages the numbers were provided by experts in the ELE consortium.[11]

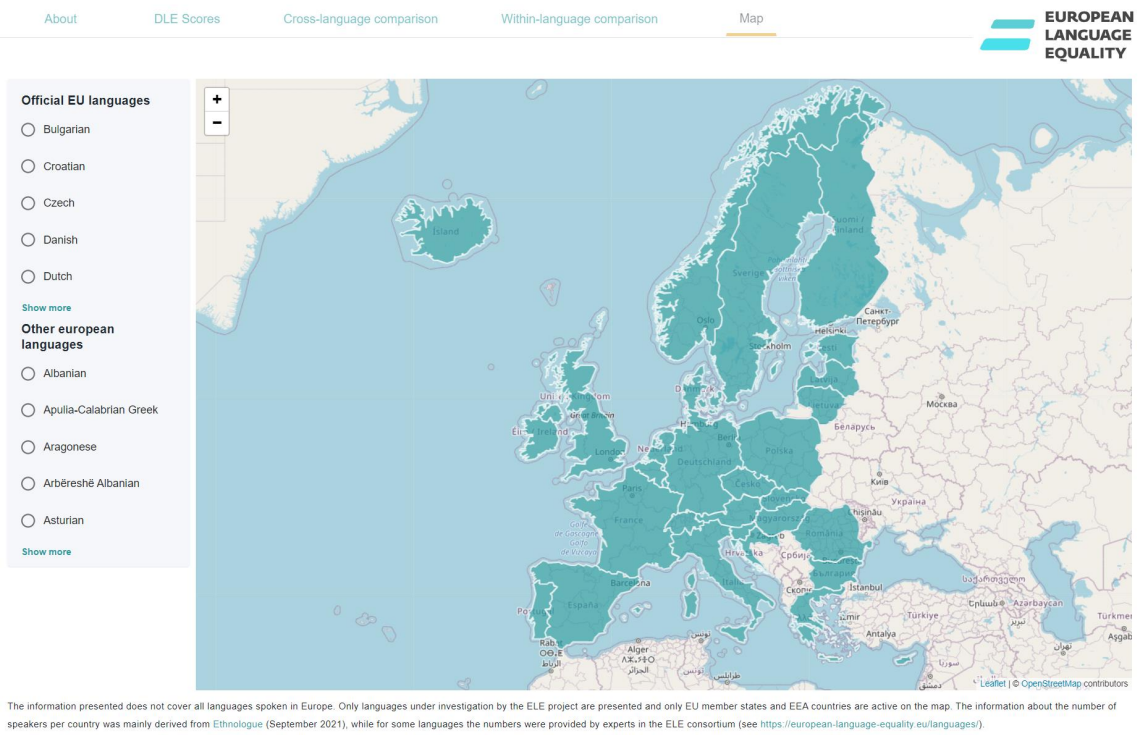

Figure 9: Dashboard: Map interface

4 Summary and Conclusions

The current ELG database encompasses to date more than 12,000 LRTs. These include (i) the ELG's own resources, (ii) the constantly growing set of metadata that are being harvested from other well-established LT-related repositories, and (iii) the metadata collected by the ELE language experts through the large-scale metadata aggregation activity for all major European languages and many of Europe's lesser spoken and/or regional languages. As such, we believe it provides an empirical basis that is as representative as possible and on which the technological readiness of the languages spoken in Europe can be measured.

The ELE dashboard presented in this report is a web user interface that exposes and visualises the level of technology support that the European languages enjoy. The dashboard allows for cross-language comparisons of the technological and contextual scores of the DLE metric and for cross- and within-language comparisons of various quantitative aspects of the ELG database, i. e. of the LRTs and some of their features that contribute to DLE.

[10] www.ethnologue.com
[11] For more details on this, see https://european-language-equality.eu/languages/

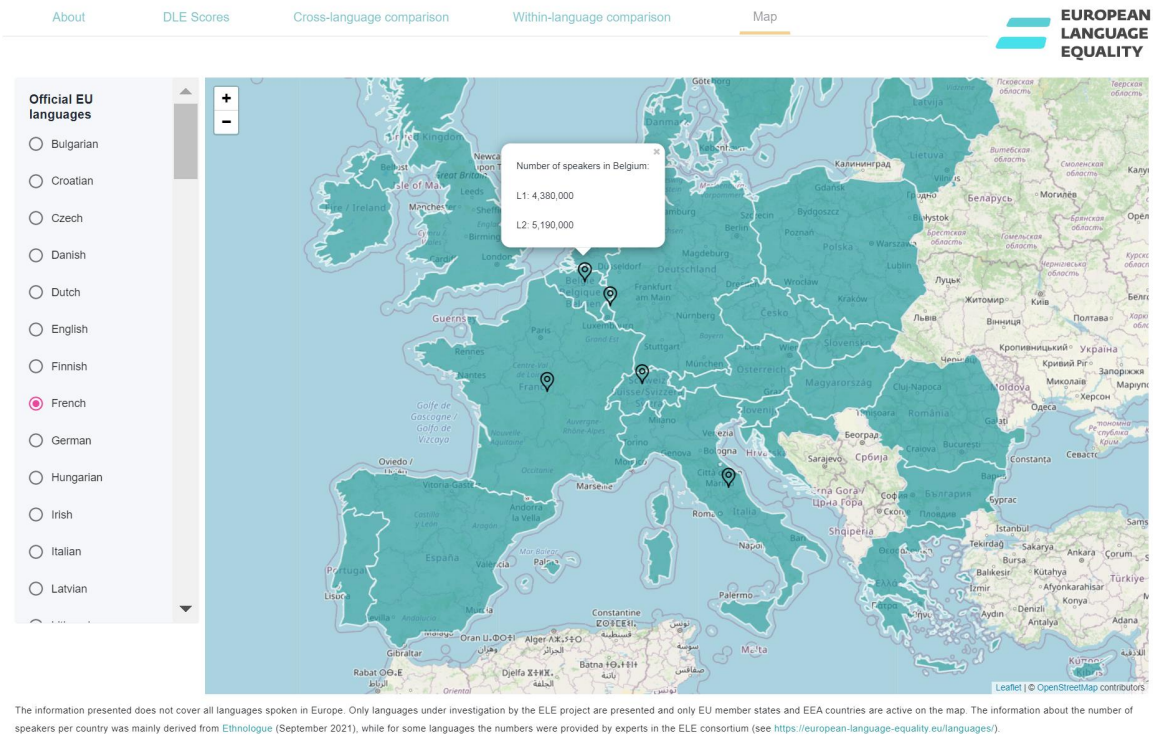

Figure 10: Dashboard: Map interface with the French language selected

The ELE dashboard facilitates the investigation of the current state of European languages with respect to their technological readiness and, going forward, it can help in the identification of potential issues or gaps for each of the languages covered. Moreover, it will help monitor the evolution of technological support, thus making the results of Task 1.3 of the ELE project sustainable beyond the task itself and the whole project. As the ELG Catalogue organically grows over time, the technological score of the DLE Metric will be updated for all European languages, thereby providing an up-to-date and consistent (i. e., comparable) measurement of the level of LT support and provision that each of them enjoys, also showing where the status is less than ideal or not at the expected level. Similarly, the situational indicators that are reflected by the contextual scores will be updated for the relevant languages by basing the computations on fresh data, as it becomes available from the selected high-quality sources.

The automatic enrichment procedures of the ELG Catalogue put in place in collaboration with the ELG project will continue at regular intervals, ensuring that the empirical basis for monitoring the level of digital readiness of Europe's languages continues to expand, reflecting future community activities and achievements. In parallel, the technical means made available through the ELG Platform will help keep the empirical basis as up-to-date as possible through easy-to-use data and metadata registration functionalities.

References

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.

Federico Gaspari, Andy Way, Jane Dunne, Georg Rehm, Stelios Piperidis, and Maria Giagkou. Deliverable D1.1 Digital Language Equality (preliminary definition), 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/05/ELE_Deliverable_D1_1.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.

Penny Labropoulou, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva. Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3421–3430, Marseille, France, 5 2020. European Language Resources Association (ELRA).

Appendix

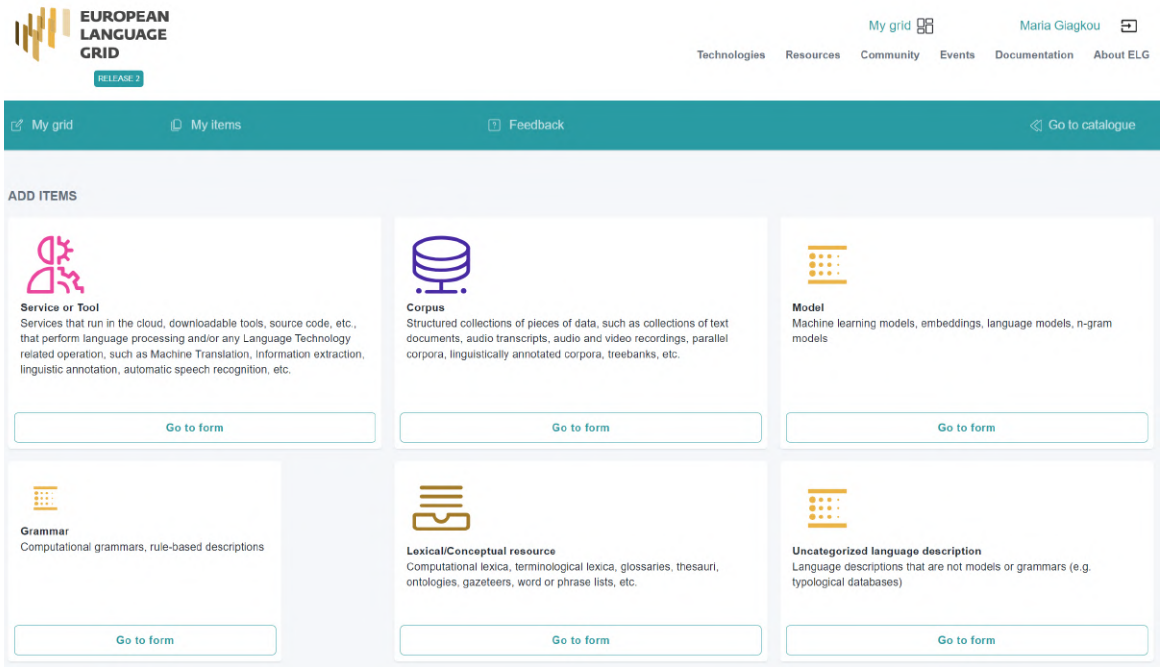

Figure 11: ELG metadata editor

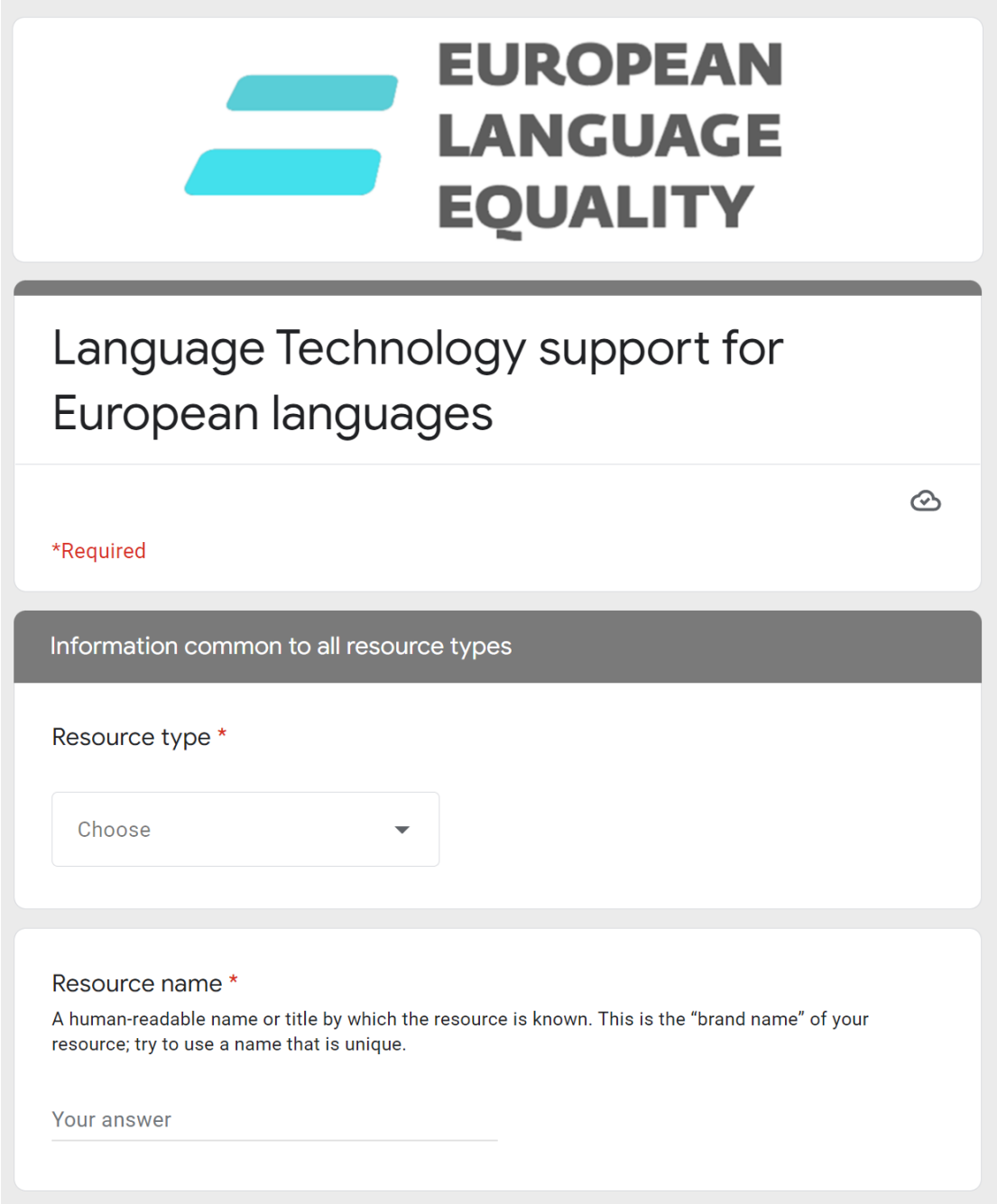

Figure 12: Online form for the documentation of identified LRs by ELE informants