EUROPEAN
LANGUAGE
EQUALITY

D1.3

Digital Language Equality (full specification)

Authors	Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, Andy Way
Dissemination level	Public
Date	28-02-2022

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	D1.3
Deliverable title	Digital Language Equality (full specification)
Type	Report
Number of pages	37
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.1 Defining Digital Language Equality
Authors	Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, Andy Way
Reviewers	German Rigau, Jan Hajič
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

AI	Artificial Intelligence
API	Application Programming Interface
CF	Contextual Factor/Factors
DLE	Digital Language Equality
DSM	Digital Single Market
EEA	European Economic Area
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
ESIF	European Structural and Investment Funds
EU	European Union
GDP	Gross Domestic Product
ICT	Information and Communication Technology
IT	Information Technology
LR	Language Resource/Resources
LRT	Language Resource/Resources and Technology/Technologies
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
NCC	National Competence Centre
NLP	Natural Language Processing
SRIA	Strategic Research and Innovation Agenda
TF	Technological Factor/Factors
TRL	Technology Readiness Level
WP	Work Package

Abstract

This deliverable provides the full operational specification of Digital Language Equality (DLE) and of its associated Metric based on a well-defined set of quantifiers, measures and indicators. Due to its descriptive, diagnostic and predictive value, the DLE Metric will be used to achieve DLE for all European languages by 2030, as a key component of the sustainable evidence-based strategic research, innovation and implementation agenda (SRIA) and of the roadmap produced by the ELE project to guide future interventions promoting Language Technologies (LTs) and language-centric Artificial Intelligence (AI) in Europe. The detailed description of DLE and of the associated Metric presented here is the result of the joint efforts of the ELE Consortium, and reflects broad consensus in the relevant communities that the ELE partners represent.

The structure of the deliverable is as follows: Section 1 introduces the background and motivation of the work presented here, and outlines its key objectives in the context of the ELE project. Section 2 presents the full specification of the DLE concept and of the DLE Metric, explaining how this work builds on the preliminary working definitions introduced in ELE deliverable D1.1. In addition, recent related studies are reviewed that were conducted on a broader scale compared to the focus of ELE on Europe, in that they cover all of the world's languages, but addressing similar concerns that are also of interest to our work, especially with regard to the immensely variable LT support enjoyed by languages, in relation to situational and contextual factors. Section 2 also emphasises the dynamic nature of the DLE Metric, and explains how it can serve as a tool to track progress towards DLE for all languages of Europe.

Section 3 presents the Technological Factors (TFs) that make up the DLE Metric, describing the central role of the European Language Grid (ELG) Catalogue as the ground truth and empirical basis to measure the level of digital readiness of the languages covered by the project. This section also discusses the scoring and weighting mechanism adopted in the DLE metric, giving an overview of the features and feature values to which weights can be assigned for the computation of the DLE Metric; the deliverable presents a first implementation of the weights assigned to the TFs. Section 3 also discusses the key decisions that were made concerning specific features of the Language Resources (LRs) and tools that make up the TFs for the purposes of the scoring and weighting mechanisms of the DLE Metric.

Section 4 presents the complementary side of the DLE Metric, the Contextual Factors (CFs), describing the data sources that were considered and eventually selected to populate the relevant indicators; part of the discussion is dedicated to the crucial and challenging issue of how country-specific data from the identified sources is apportioned to the respective languages and language communities for the purposes of quantifying the CFs of the DLE Metric. The relevant data preparation process is outlined, laying emphasis on the issues of quality of the data and on the possibility of refreshing the data underlying the CFs with regular updates, in order to support the dynamic nature of the DLE Metric. On this basis, the process of computing the DLE Metric scores for CFs is illustrated with detailed examples, considering the merits of alternative configurations of the CF set-up. These were assessed as part of a heuristic evaluation conducted by experts, who provided feedback in order to select the most convincing and effective mix of CFs for the DLE Metric in relation to the specific features of the languages and language communities under consideration. The end of Section 4 reviews a number of suggestions and opportunities for future improvements to the CFs of the DLE Metric that were proposed as part of this heuristic expert evaluation.

Finally, Section 5 draws some overall conclusions, discussing the central role of the DLE Metric in realising the vision of DLE for all languages of Europe by 2030, and emphasising its relevance for the LT and language-centric AI community as well as for a wide range of stakeholders, ultimately for the benefit of all European citizens.

1. Introduction

1.1. Background and Motivation

This document builds on ELE deliverable D1.1 *Digital Language Equality (preliminary definition)*, extending the foundational work presented there in several important directions. The preliminary working definitions of the concept of Digital Language Equality (DLE) and of the DLE Metric proposed in D1.1 have guided the efforts of the ELE project and its Consortium so far. Since then, these foundational concepts have been gradually refined and focused to achieve an operational full specification of DLE to drive the key tasks of the remaining part of the project and of its planned developments.

The progress documented in this deliverable was led by the Core Group and sustained by the joint work of the ELE Consortium, which has had several opportunities to present and discuss the concept of DLE and the associated Metric with a wide range of external stakeholders in various forums. This community-driven endeavour has enabled the collection of valuable feedback and additional input to define the details of DLE, so that it can have a powerful descriptive, diagnostic and predictive value to successfully promote full digital equality for all the languages of Europe. In this spirit, the work presented in this deliverable has benefited substantially from the close collaboration with its sister project, the European Language Grid (ELG),[1] which continues to provide solid empirical evidence and ground truth on which the computation of the DLE Metric is based, as explained in detail in Section 3.1.

1.2. Key Objectives

The updated specification of DLE and of the associated Metric presented in this deliverable will support the sustainable evidence-based strategic research, innovation and implementation agenda (SRIA) and the roadmap for achieving full DLE in Europe by 2030 that will be produced by the ELE project to guide the ELE Programme in the coming years. Accordingly, this work will contribute to ongoing efforts to avoid the danger of extinction that is still impending for some languages spoken in Europe, and will drive concerted actions to level up Language Technology (LT) support for each and every language of Europe, so that they can all "continue to exist and to prosper as living languages in the digital age", as stated in the preliminary definition of DLE included in D1.1. To this end, the aim of this deliverable is to fully articulate the definition of the DLE Metric, by detailing the specifics of the Technological and Contextual Factors (TFs and CFs, respectively) that contribute to the computation of the DLE Metric score for all the languages covered by the ELE project.

The existence of a carefully designed, empirically grounded and widely agreed full specification of DLE and its associated Metric will be instrumental in prioritising and implementing interventions to raise the level of LT support across Europe. ELE is in the unique position of representing a comprehensive cross-section of the European LT community, and thanks to the close connection with the ELG project, it is also able to draw on the ELG Catalogue as the cornerstone of its empirically-driven SRIA and roadmap. Based on the work presented here, ELE will produce a Dashboard (to be presented in D1.35) to interactively visualise the indicators of the level of LT support for the languages covered by the project. This concerted effort provides an unprecedented opportunity to the European LT and language-centric AI community to guide future targeted funding programmes that can effectively leverage national, regional and local resources in synergy with EU-wide schemes, to secure the much needed mix of funding that can support the achievement of DLE for all of Europe's languages by 2030, with substantial benefits for all European citizens, society and economy, including at Member State level.

[1] https://european-language-grid.eu

2. Full Specification of Digital Language Equality

2.1. From the Preliminary Definition to the Full Specification of Digital Language Equality and its Metric

As stated in its preliminary definition expressed in D1.1, the DLE Metric “is a measure that reflects the digital readiness of a language and its contribution to the state of technology-enabled multilingualism, tracking its progress towards the goal of DLE”. The initial definition also stipulated that both the level of “technological support” and the “situational context” of the languages should play a role in DLE, as two complementary dimensions of language equality in the European context. The ELE Consortium has built on the work presented in D1.1 to formulate the full operational specification of DLE and its associated Metric that are introduced here, after availing of several opportunities to take on board the views of players in the broader European LT and language-centric AI community, thanks to the extensive network of contacts and collaborations of its members. The fully articulated detailed definition presented here will guide the tasks and actions of the remaining period of the project, especially for the development of the SRIA and roadmap that will provide the detailed plan to achieve full DLE in Europe by 2030.

While any quantification of any measure is bound to be subjective and arbitrary, at least to some extent, the full specification of DLE endorsed by the ELE project enjoys unprecedented and unparalleled support in the community, insofar as it is considered applicable to all languages of Europe and, by extension, relevant to all the countries, regions, areas and language communities where these are spoken. Crucially, the fully articulated definition should not only have descriptive value, i. e., be able to accurately reflect where all European languages currently stand in the legitimate aspiration to achieve full DLE, but also – and more ambitiously – serve diagnostic and predictive purposes; these involve a dynamic and forward-looking function of the DLE Metric (Section 2.3), i. e., the scores assigned to each European language will specifically indicate over time which ones lag behind and are in need of interventions, and show evidence of how they eventually benefit from the enhanced provision of LRs, tools and technologies, as dedicated funding is released.

This effort is targeted at all of Europe’s languages, recognising their specific realities and different roles in the respective communities of speakers, to ensure that they are all well-equipped for the future, to meet the challenges of the digital era. Following in the tradition of the META-NET White Paper Series (Rehm and Uszkoreit, 2012) a decade on, ELE Deliverables D1.4-D1.34, D1.36 and D1.39 provide an update of the overall current status of technological support for Europe’s languages, on the basis of a wealth of solid empirical evidence and expert opinion. Crucially, these deliverables reveal that for some European languages the prospect of extinction is still a very real one, and that most of them can still rely only on a limited level of technological support, that ranges from fragmentary, to weak, to (virtually) non-existent. This situation requires prompt targeted action to redress the balance of DLE, and the Metric presented here is the tool developed by the community to achieve this pressing objective for Europe and all of its people, recognising the central value of languages in their cultural heritage and distinctive identities.

It is important to note that the DLE Metric can serve two complementary functions for all of Europe’s languages, namely intra- and inter-linguistic. This means that DLE Metric scores measured at regular intervals can first of all objectively track the progress of LT support in the interest of DLE of each language, to monitor its progress in the desired direction over time, thanks to the dynamic and updatable nature of the Metric, as explained in more detail in Section 2.3. Secondly, DLE Metric scores also provide an additional comparative measure of the relative positioning and improvements over time of different languages, for instance (but not exclusively) those that have similar current levels of LT support (Section 3) and/or comparable CFs (Section 4), or to examine the development of neighbouring languages, e. g. with a view to evaluating the effectiveness of targeted national, regional and local funding interventions to enhance their overall LT provision.

2.2. Recent Related Work

In addition to relying on the input and advice of Europe's LT and language-centric AI community, the work conducted in ELE to formulate the full specification of DLE and its associated Metric has also drawn inspiration from recent related efforts with a broader scope, that extends beyond the European scenario, but were deemed relevant due to substantial common ground. Particularly interesting and relevant examples of such related work with a global coverage are described in Joshi et al. (2020), Blasi et al. (2021) and Bromham et al. (2021): in what follows we discuss these recent prominent contributions, highlighting the main points of contact and shared concerns with our work in ELE.

First of all, Joshi et al. (2020) cover the over 7,000 languages of the world to investigate the relation between the types of languages, available resources and their coverage in Natural Language Processing (NLP) conferences. Their objective is to trace the evolution of the attention devoted by the international NLP community as a whole to the languages of the world, providing evidence for the severe disparity that exists across languages in the scale of attention and technological support, as estimated on the basis of the number of mentions that languages receive in NLP conference papers over time. As a result of this wide-ranging global analysis, Joshi et al. (2020) propose a taxonomy that groups the languages of the world featured in leading NLP conference proceedings into 6 categories in the quest for technological support, which they define "the left-behinds, the scraping-bys, the hopefuls, the rising stars, the underdogs and the winners". Interestingly, but not surprisingly, at the opposite extremes of this spectrum, the most disadvantaged category of left-behinds alone accounts for well over 2,000 languages, for a combined total of 1.2 billion speakers across the globe; at the opposite end, the two most privileged categories (i. e., underdogs and winners) combined include just 25 major languages in total between them, all used primarily in developed and advanced countries.

Secondly, considering similar issues, Blasi et al. (2021) argue that the substantial progress brought about by the generally improved performance of NLP methods "has been restricted to a minuscule subset of the world's 6,500 languages", and present a framework for gauging the global utility of LTs in relation to demand based on the analysis of a sample of over 60,000 papers from all major international NLP conferences. They also show convincing evidence for the "immense inequality in the development of language technologies across the world's languages. After English, a handful of Western European languages dominate the field": in other words, major efforts in global NLP work are predictably focused on an extremely small set of elite languages that enjoy privileged socio-economic positions on the international scene. While this severe imbalance is in favour of a few, mostly European, languages, on the whole the situation is very uneven, and most other European languages are themselves at a disadvantage. The work of Blasi et al. (2021) was supported by the US-based National Science Foundation, and interestingly it also discusses some of the key (latent) societal, economic and academic factors that cause, and at the same time fatally reinforce, the blatant disparities that they identified. By way of conclusion, they propose a set of specific recommendations to encourage evidence-based policies "aimed at promoting more global and equitable language technologies", with a focus on academic and industrial research.

The third particularly relevant related study reviewed here due to its strong connections with our work in ELE, albeit on a broader global scale, is that by Bromham et al. (2021). They analysed 6,511 spoken languages of the world (which corresponds to over 90% of the total, according to most statistics) according to 51 predictor variables of language maintenance concerning various aspects having to do with speakers' population, documentation and legal recognition of the language, education policy, socio-economic factors and environmental conditions. In addition to emphatically arguing that "language diversity is under threat" across the world, Bromham et al. (2021) point out in particular that a greater development of transport infrastructure, and especially road density, in a region are linked to increased local language endangerment: by encouraging the circulation of the population, such external, non-linguistic conditions undermine the role and preservation of languages already under pressure. Another socio-educational factor that Bromham et al. (2021) found to decisively contribute to language endangerment concerns the longer periods of formal education received by the youngsters of a language community in an official language, which may lead to a higher likelihood of not preserving the heritage language in actual active use into adult life.

In particular, the research found that 37% of the world's 6,511 languages under investigation are considered to be threatened or endangered (i. e., losing first-language speakers or only spoken by adults, without child learners), while 13% were placed in the even less enviable category of "sleeping" (i. e., no longer spoken as first languages): overall, this means that around 50% of the investigated languages (i. e., over 3,000 of them across the world) face serious risks of extinction, potentially within a generation, if not imminently. The sombre and ominous conclusion of this very wide-ranging and solid study is that "[w]ithout intervention, language loss could triple within 40 years, with at least one language lost per month". While it is unclear whether this general pattern also applies specifically, with equal devastating force, to the less-resourced languages of Europe, this is certainly a sobering global reality to face up to, which calls for a large-scale mobilisation of all possible efforts by all interested parties to avoid such a daunting prospect propagating to the languages addressed by ELE.

This review of recent relevant work is supplemented by a fourth timely and interesting study, which focuses specifically on the non-linguistic factors that affect the dynamics at play in language communities. Faisal et al. (2021) investigate geographical and economic factors, determining the origin of data sets using an entity recognition method. The respective predictive values are calculated for these CFs in order to examine the socio-economic correlations in connection with the distribution of the data sets (Faisal et al., 2021). Three factors were investigated: the gross domestic product (GDP), size of the language community, and geographic proximity. Most of the data sets came from countries considered to be economically prosperous, so the best predictive value was GDP, and the best results occurred when taking GDP and geographic proximity into account. Thus, this interesting study shows that an interaction of several factors seems to improve prediction.

Interestingly, these well-researched recent studies with an inclusive interest in the world's languages establish clear links between the level of technological support and development on the one hand, and a variety of situational conditions on the other, in order to comparatively analyse the level of LT support for a very broad and diverse range of languages and the potential for their communities to see their needs and aspirations served in the digital era. Otherwise, if this were not to be the case, the very clear and concerning prospect is that many of the languages that are not adequately supported by technology appear to be doomed to oblivion in the digital age, with very limited possibilities even for dignified survival. In addition to building on the preliminary work presented in D1.1, this deliverable also draws from Joshi et al. (2020), Blasi et al. (2021), Bromham et al. (2021) and Faisal et al. (2021) as inspiring and thought-provoking studies with a clear relevance for ELE. As a result, in addition to the TFs discussed in Section 3, in developing the full specification of DLE and its associated Metric, the partners of the ELE project endeavoured to also take into account the similarly crucial situational conditions that seem to play such a determining role in securing or jeopardising the future of languages in an increasingly interconnected world dominated by technology; such CFs that play a role in the DLE Metric are presented in Section 4.

2.3. Dynamic Nature and Updates of the DLE Metric

A crucial feature of the DLE Metric is its dynamic nature, i. e., the fact that its scores can be updated and monitored over time, at regular intervals or whenever one wishes to track the progress or the status of one or more European languages with regard to the goal of achieving DLE. This dynamic nature involves the two related aspects of TFs and CFs, and this time-sensitive nature of the DLE Metric will be supported in the ELE Dashboard that will be introduced in D1.35. With regard to the TFs, as the ELG Catalogue organically grows over time (see Section 3.1), on the basis of the weights that are assigned to the feature values (Appendix A), the resulting DLE Metric scores 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 CFs will be updated for the relevant languages by basing the computations of the DLE Metric visualised through the ELE Dashboard on fresh data, as it becomes available from the selected sources (Section 4.1).

3. Technological Factors

3.1. The ELG Catalogue as the Ground Truth at the Basis of the Technological Factors

The full specification of the TFs included in the DLE Metric articulated in this deliverable was devised by analysing the actual contents of the ELG Catalogue for all European languages between the end of 2021 and late January 2022, as the empirical evidence that provided the firm basis on which the final structure of the DLE Metric was eventually defined; at that stage, in late January 2022, the ELG Catalogue contained approximately 11,500 records, out of which about 75% were language data resources (corpora, lexical resources, models and grammars) and the rest were LTs (tools/services).

A significant number of LRs and tools for all the languages of Europe in the scope of the ELE project were identified by the relevant experts and informants, and described with the relevant metadata that are essential for the subsequent computation of the DLE Metric. The language experts had full control, ownership as well as responsibility for identifying and documenting LRs for their languages, and they were ideally positioned to perform this essential role, due to their being in most cases ELG National Competence Centre (NCC) leads and European Language Resource Coordination (ELRC) National Anchor Points,[2] hence fully aware of the current LR and LT provision for their languages in the broader landscape, and therefore also cognisant of what is required to boost DLE across Europe, specifically targeting individual languages. The metadata of the identified LRs were subsequently processed, normalised and imported into the ELG Catalogue, further expanding its own collection of LRs. By examining the actual richness and breadth of the metadata associated with the ELG Catalogue records with the support of representatives of the LT and language-centric AI community, we were able to establish the relevant features and feature values that had to be included in the final specification for the computation of the DLE Metric, and the results of this work are reported in this document.[3]

[2] https://www.european-language-grid.eu/ncc/ and https://lr-coordination.eu/anchor-points

[3] The first proposed implementation of the DLE Metric presented in Appendix A does not take into account the size of LRs and the quality of LRs and tools. While these are important features, there exist a large variety of size units for LRs, and the way for measuring data size is not standardised, especially for new types of LRs such as models. Regarding the quality of tools in particular, while some information on the Technology Readiness Level (TRL) is available, the large number of null values does not make it possible to take this into account at

3.2. Scoring and Weighting Mechanism of the DLE Metric

One guiding consideration in developing the DLE Metric, and especially in assigning the weights of the features and their values for the TFs, was that we did not want to make assumptions about the possible (preferred) end-uses and actual application scenarios that may be most relevant to users. These inevitably vary widely due to a number of variables that are impossible to establish a priori. We therefore refrained from predetermining at this stage particular preferred end-uses when proposing the full specification of the DLE Metric, which otherwise would risk it being unsuitable for some end-users and applications. As explained in more detail in Section 3.3, here we present the DLE Metric with a first proposed set of weights for the CFs in Appendix A, subject to revision as more experiments are run within the ELE project to adjust the weights, so that the DLE Metric scores capture and reflect fairly the actual level of LT support for the ELE languages. This work will feed the ongoing implementation of the ELE Dashboard, that will be presented in D1.35.

Appendix A presents in detail the features and associated values for LRs and tools that make up the TFs, which are derived from the ELG Catalogue metadata schema (as described in Labropoulou et al. (2020) and Rehm et al. (2020), and explained in detail in D1.1), with a first proposed assignment of weights. Here we briefly review some of the key features of the TFs, focusing in particular on those that can have several values which are of particular interest, e. g. insofar as they show the level of detail and granularity of the metadata accompanying the records included in the ELG Catalogue. This is the case, for instance, for the "Subclass" feature within LRs, that can have a range of as many as 23 values: apart from "raw corpus", to which we have assigned a nominal minimal weight, those that were deemed to be worthy of particularly generous weights in the first implementation presented in Appendix A are (in descending order) "model", "Wordnet" and "Framenet", "terminological resource", "annotated corpus" and "morphological lexicon"; in addition, there are 15 other possible Subclass values, that were rewarded comparatively less, with a very moderate fixed and constant weight.

A particularly rich feature within LRs is that of "Annotation Type", which has many possible values. For the first implementation proposed in Appendix A, we have assigned a constant very small fixed weight, also based on the fact that some LRs can possess several annotation types. A similar consideration applies to the "Domain" feature, which has very many possible values for LRs and for tools: in these cases, the weights assigned to "Domain" values in the first instance are fixed and relatively small, again considering that multiple domains can be combined in a single LR or tool. It is important to note that for the features "Annotation Type" (in LRs) and "Domain" (both in LRs and tools), we allow for the possibility to assign flat scores that are identical to all of the many potential values, that may be difficult to differentiate in abstract terms across the board for all of Europe's languages. In addition to "Domain", another feature that appears both in LRs and tools is "Conditions of use"; the initial weights proposed in Appendix A for this feature of the TFs are identical for the corresponding values of "Conditions of use" across LRs and tools. In the case of (much) more restrictive licensing terms, lower weights are assigned than to liberal use conditions, hence they contribute (much) less to the DLE score for the LR in question, and therefore to the cumulative DLE Metric score for that language.

The proposed specification of the TFs and the relevant weights for the feature values to be used for the computation of the DLE Metric were initially discussed at length within the ELE Core Group, and then presented at a General Assembly of the ELE project to all Consortium partners in early February 2022, to gather their feedback and input on how to finalise

the moment. This is an area well worth revisiting in subsequent efforts to extend the DLE Metric with regard to the TFs. Similarly, no specific weights have been proposed for projects and organisations for the time being, partly due to the difficulty of attributing them specifically to individual languages, even though these additional features may be included in the DLE Metric at a later stage.

the set-up, so that the computation of the TFs could enjoy their full support to drive the major remaining tasks of the project. This was a very valuable opportunity to refine elements that were not perceived to be completely adequate for some of the languages targeted by the project, and ensure that the final full specification of the DLE Metric could be applied across the board to all of Europe's languages: this deliverable describes the results of this inclusive and comprehensive approach to the definition of the DLE Metric, and the initial implementation of the TFs is presented in Appendix A. The required adjustments to the weights of the feature values in the various internal experiments with different potential weighting set-ups for the TFs were made by developing a preliminary offline basic database that could replicate the essential features of the proper ELE Dashboard that is currently under development and is due to be presented in D1.35. This mock-up used to adjust the weights assigned to the feature values of the TFs was developed in the form of an Excel spreadsheet file containing the ELG Catalogue export as of late January 2022.

This dataset included metadata of both LRs and tools for all ELE languages. Each resource and tool has several features and associated values, as shown in Appendix A. Each of these features was then given a weight to calculate the DLE Metric LR score, the DLE Metric Tool score, and the total combined DLE Metric score on a per language basis. When experimenting with the preliminary settings of the weights for the DLE Metric, users from the ELE Core Group applied weights of their own choice in the "LR Factors" and "Tools Factors" sheets in the Excel file. The results of these weights could be seen in a "Language Scores" sheet, which also showed the aforementioned scores for each language. The scores were also represented as bar charts, a very simple but intuitive visualisation of where each language stood relative to all the others. Any changes to the weights were automatically implemented, updating the scores, and subsequently updating the bar charts. This was an efficient and effective method to gradually refine the set-up of the TFs and propose the first implementation of the relevant weights presented in Appendix A.

On the whole, the experiments conducted in preparation for this deliverable with different weights and set-ups of the DLE Metric have shown that the global picture of the DLE Metric scores for the languages targeted by the project is unlikely to change dramatically. As a matter of fact, we have experimented both with very moderate and narrow ranges of scores and weights assigned to the various features and their values of the TFs, and with more extreme and differentiated weighting schemes. Since, ultimately, any changes are applied across the board to all LRs and tools included in the ELG Catalogue for all languages, any resulting changes propagate proportionally to the entire sample of languages, thus making any dramatic changes rather unlikely, unless one studiously unduly rewards specific features that are known to disproportionately affect one or more particular languages. It should immediately be clear that this would be a biased and unfair application of the DLE Metric, and should be avoided at all costs.

In any case, our experience in experimenting with the weights of the DLE Metric has shown that artificially inflating or diminishing the scores of specific languages, especially with regard to other similarly positioned languages, is quite difficult and on the whole unlikely to happen. In essence, after experimenting with various DLE Metric set-ups and weights for the TFs, we can conclude that the resulting representation tends to be relatively stable. This is due partly to the sheer amount of features and possible feature values that make up the TFs, which are grounded in the metadata of the ELG records, as discussed in Section 3.1, but to some extent also to the different level of overall provision of LTs that each European language is equipped with. As a result of this, even if one tweaks the weights assigned to TFs, with the exception of relative minor and local fluctuations, three main phenomena are generally observed: (i) the overall relative positioning of the languages remains largely stable, with a handful of languages standing out with the highest DLE Metric scores (English leading typically over German, Spanish and French, with the second language having roughly half the DLE Metric score as English), and the minimally supported languages still displaying very low scores, and a substantial group of evenly distributed languages in the middle; (ii) clusters of languages with similar LT support that is relevant to ensure DLE in Europe will remain ranked closely together, regardless of the adjustments made to specific weights for individual features and their values; and, finally, (iii) even when two similarly supported languages change relative positions (i. e., language A overtakes language B in terms of DLE Metric score) as a result of adjusting the weights assigned to features and their values, their absolute DLE Metric scores remain very close, thus not substantially distorting the overall representation of their actual status of LT support.

In addition to investigating the effects of changing the set-up of weights for the TFs across all ELE languages as a whole, we have also performed focused checks on pairs or small sets of languages that operate in similar circumstances, and whose relative status in terms of LT support is well known to the relevant experts. These focused checks have involved, for example, Basque and Galician, Irish with respect to Welsh, and the dozen local languages of Italy (also with respect to Italian itself), etc. Overall, the general stability demonstrated by the DLE Metric across different set-ups of weight assignments for the various features and their possible values for TFs provides evidence of its validity as an effective tool to guide developments and track progress towards full DLE for all of Europe's languages by 2030.

3.3. Flexibility of the DLE Metric Weights for the Technological Factors

Against this background, Appendix A provides a first possible configuration of the weights that can be assigned to the TFs, which is the result of current consensus within the ELE Consortium and has been finalised following consultation at the General Assembly in early February 2022. This first configuration of the weights is provided in Appendix A, subject to adjustments as more experiments are conducted to check any needs to tweak the weights, in the interest of making the DLE Metric truly representative of the level of LT support that European languages rely on. This approach will ensure that the Dashboard that will be presented in D1.35 represents a suitable tool to optimally capture the current situation and also effectively reflect the needs and aspirations of all of Europe's languages for the future in the digital age.

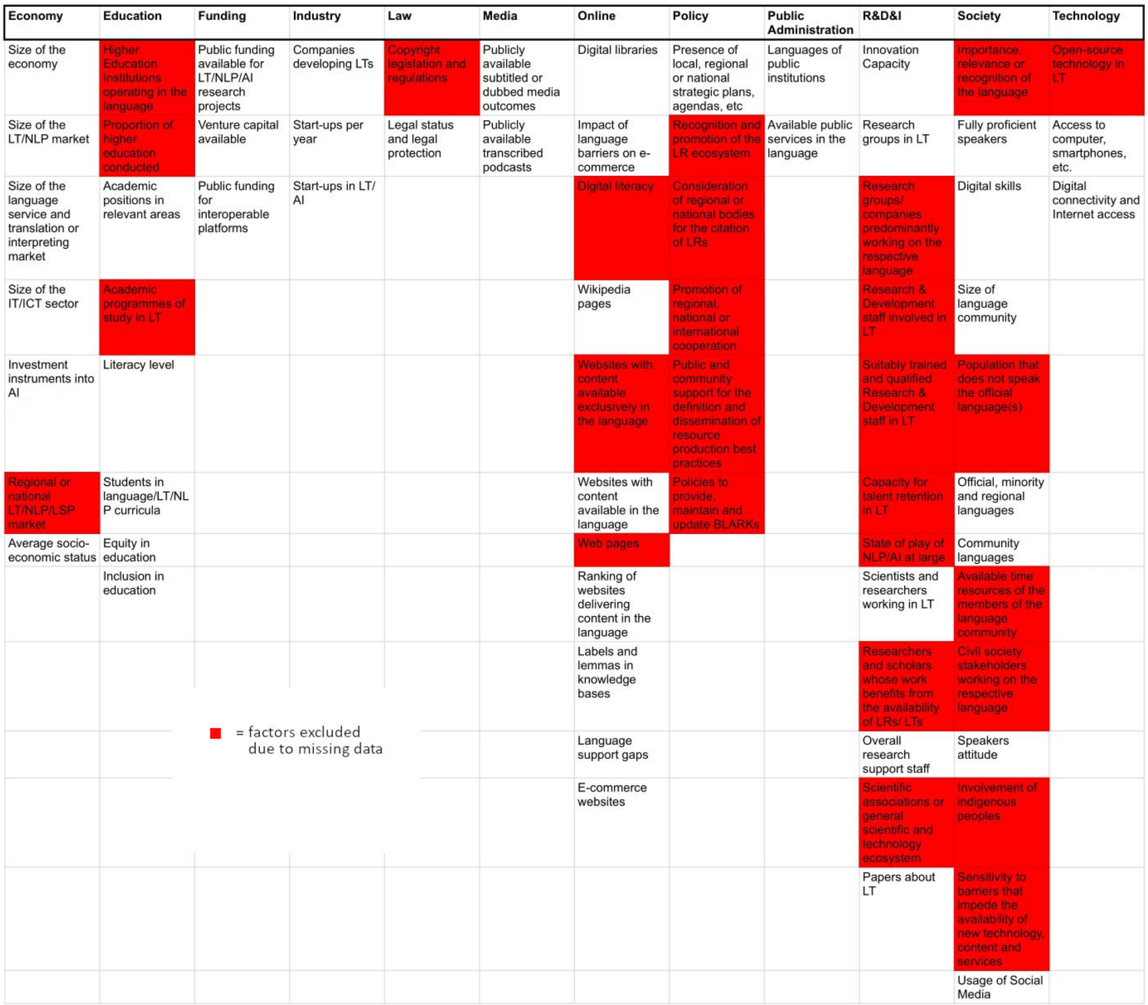
In this spirit, the ELE Consortium wishes to remain open to the possibility of adjusting the parameters, including the features and the relevant values and the exact configuration of the weighting scheme adopted for the TFs, which could be done at regular intervals on the basis of agreement among the relevant parties. This approach is useful to address developments ensuing as a result of advances being made in LT and as new paradigms or technologies become the state of the art, potentially also as new types of resources emerge and are recognised as crucial for LT support. In other words, we view the DLE Metric as a flexible tool, that can allow for changes and updates at future points in time to the features of the TFs and to the weights assigned to the relevant values, as a result of how the state of the art evolves, with the DLE Metric changing accordingly.

4. Contextual Factors

4.1. Data Sources

In the preliminary definition of DLE presented in D1.1, 72 contextual factors were presented and clustered into 12 classes representing different aspects of the context (Gaspari et al., 2021). Each of the factors had to be quantified with an indicator so as to be measurable. The quantification depended on the presence and accessibility of data for an indicator being feasible to represent the factor. Therefore, different data sources with pan-European data were collected. The selected sources included, among others, EUROSTAT, the European Language Monitor, Glottolog and reports or articles. A list of the factors, the chosen indicator and the relevant data source can be found in Table 3 in Appendix B.

Overall, 26 factors were excluded due to missing data. This affected especially factors from the classes “research & development & innovation”, “society” and “policy”. Data about policies is mainly too broad and just represents whether policies exist or not. The class “society” included factors about diversity being difficult to quantify. The problem of missing data in this area was already mentioned in the AI Index report (Zhang et al., 2021). The factors excluded from the class “research & development & innovation” covered mainly specific figures about the research environment of LTs, while broader figures about the research situation of the whole country independent of research areas are available. Figure 1 shows all factors presented in the preliminary definition. The cells highlighted in red were eventually excluded. Overall, 46 factors were quantified with an appropriate indicator.



Economy	Education	Funding	Industry	Law	Media	Online	Policy	Public Administration	R&D&I	Society	Technology
Size of the economy	🟥 Higher Education Institutions operating in the language	Public funding available for LT/NLP/AI research projects	Companies developing LTs	🟥 Copyright legislation and regulations	Publicly available subtitled or dubbed media outcomes	Digital libraries	Presence of local, regional or national strategic plans, agendas, etc	Languages of public institutions	Innovation Capacity	🟥 Importance, relevance or recognition of the language	🟥 Open-source technology in LT
Size of the LT/NLP market	🟥 Proportion of higher education conducted	Venture capital available	Start-ups per year	Legal status and legal protection	Publicly available transcribed podcasts	Impact of language barriers on e-commerce	🟥 Recognition and promotion of the LR ecosystem	Available public services in the language	Research groups in LT	Fully proficient speakers	Access to computer, smartphones, etc.
Size of the language service and translation or interpreting market	Academic positions in relevant areas	Public funding for interoperable platforms	Start-ups in LT/AI			🟥 Digital literacy	🟥 Consideration of regional or national bodies for the citation of LRs		🟥 Research groups/ companies predominantly working on the respective language	Digital skills	Digital connectivity and Internet access
Size of the IT/ICT sector	🟥 Academic programmes of study in LT					Wikipedia pages	🟥 Promotion of regional, national or international cooperation		🟥 Research & Development staff involved in LT	Size of language community	
Investment instruments into AI	Literacy level					🟥 Websites with content available exclusively in the language	🟥 Public and community support for the definition and dissemination of resource production best practices		🟥 Suitably trained and qualified Research & Development staff in LT	🟥 Population that does not speak the official language(s)	
🟥 Regional or national LT/NLP/LSP market	Students in language/LT/NLP curricula					Websites with content available in the language	🟥 Policies to provide, maintain and update BLARKs		🟥 Capacity for talent retention in LT	Official, minority and regional languages	
Average socio-economic status	Equity in education					🟥 Web pages			🟥 State of play of NLP/AI at large	Community languages	
	Inclusion in education					Ranking of websites delivering content in the language			Scientists and researchers working in LT	🟥 Available time resources of the members of the language community	
						Labels and lemmas in knowledge bases			🟥 Researchers and scholars whose work benefits from the availability of LRs/ LTs	🟥 Civil society stakeholders working on the respective language	
						Language support gaps			Overall research support staff	Speakers attitude	
						E-commerce websites			🟥 Scientific associations or general scientific and technology ecosystem	🟥 Involvement of indigenous peoples	
									Papers about LT	🟥 Sensitivity to barriers that impede the availability of new technology, content and services	
										Usage of Social Media	

🟥 = factors excluded due to missing data

Figure 1: Overview of the Contextual Factors

4.2. Data Preparation

One main attribute of the collected data that represented a challenge for the purposes of the work involved in defining the CFs for inclusion in the DLE Metric was their heterogeneity. It had varying formats, was based on country or language level, included differing languages or countries and consisted of three data types. As a result, the data had to be standardised before it could be processed. Firstly, the formats of the figures were converted into a consistent schema for numbers, namely xx,xxx.xx, as opposed to different usage of commas and decimal points. Secondly, some indicators based on the data from Eurostat had to be prepared because the data was split in different tables. For instance, the factor "researchers in LT" was quantified with the total number of researchers in the research areas "computer and information science", "linguistics and literature", "media and communication" and "humanities". Eurostat provides a separate table for each of these research areas. Thus, the numbers were added up for each country to obtain one figure per country. Moreover, the data based on the level of languages differed in the language names. Varying names like Griko and Apulia-Calabrian Greek had to be mapped with the help of Glottolog,[4] containing in each metadata record per language a list of alternative names.

Additionally, some sources consisted of plain text from which scores had to be extracted. One example is the extracted information from the website of LT-innovate[5] about the existing funding on national and European level. Three paragraphs described the funding situation regarding National/Regional Funding, European Structural and Investment Funds (ESIF) and funding through Eureka/Eurostars per country. If the text indicates the country to have funding on regional, national or international level or through ESIF or Eureka/Eurostars, then a score of 1 was assigned to each funding opportunity. The highest score is 6 representing a funding on regional, national and international level and through ESIF, Eureka and Eurostars. A list of the indicators transformed from plain text into scores and an explanation of the process is given in Table 4 in Appendix B.

Because the metric is intended to process data on a language-by-language basis, data collected at the country level had to be converted to the language level. In total, the factors were quantified with three different types of data, total numbers, proportional numbers, and scores.

Most total numbers were split proportionally, using the percentage of speakers of the language per country. The figures for the percentages were calculated through the population size and the number of speakers from Ethnologue. A collection of the figures about the number of speakers in the European countries covered by the project was done in May 2021 within the project.[6] Due to some gaps and old data records regarding regional and minority languages, experts on minority languages within the ELE consortium were asked to fill the gaps or to provide data more representative of the actual status quo. The figures for the languages Alsatian, Faroese, Gallo, Icelandic, Macedonian, Meskhetian and the Saami languages were corrected. Based on this data, the calculation of the proportion of the language community was done.

Percentages of languages preferentially taught in education and thus often second languages (particularly English, German, French, and Spanish) were only included if the language had an official status in the country. For example, the figures of English are based on the figures of the UK, Ireland and Malta. In the other European countries, English does not have an official status, so they are not taken into account. If the language was an official language in another country, only language communities with a higher percentage than one percent were covered to simplify the mapping. This calculation was performed for each language community in each of the European countries covered by the project.

[4] https://glottolog.org
[5] https://www.lt-innovate.org/lt-observe/public-policy-observatory/national-funding-opportunities
[6] https://european-language-equality.eu/languages/

Total numbers per capita, proportional numbers, and scores were applied to the language communities without adjustment due to the complexity and additional time the adaption would have needed. A complex mapping would be desirable, as many language communities – especially members from minority language communities – deviate from the average. Additionally, the mapping via the proportion of the speakers is also somewhat problematic, since in some cases the sum of the speaker communities is not 100%. For example, numbers from countries with many bilingual speakers were given several times to the different language communities. If the calculation were reversed, these countries would have a higher GDP than they actually have. Another problem is the missing inclusion of the political reality regarding the protection or promotion of a language. This refers in particular to figures like how many researchers work on the language, which were also transferred to the languages by a percentage mapping. In countries with a large number of speakers of a language, but where less funding or research activity is devoted to the preservation and promotion of that language, a direct mapping does not fit. For example, in Ireland, little promotion and funding is invested in the Irish language compared to the second national language, English. The number of speakers in largely private contexts is relatively high, at 39.8 percent according to Wikipedia[7]. The numbers of factors such as "scientists and researchers working on LT and the respective language" are relatively high due to the calculation of the percentage, although in reality only a small fraction of researchers in LT work on the Irish language. Figures on the number of companies in the ICT field are also very high for Ireland due to the low tax rate for large companies. Many companies based in Europe therefore locate in Ireland but do not support development for Ireland and even less for the Irish language community, but operate their European businesses from there. This specific environmental, political, and social circumstances are not considered in a percentage-based mapping.

If a language was spoken in more than one country, the total numbers were added up, but proportional numbers, scores and total numbers per capita were calculated using the average. At this point the different sizes of the language communities were taken into account, so the data values of bigger language communities were weighted double at the calculation of the average, meaning the number of the bigger language communities were taken twice into account.

4.3. Calculation of the Contextual Factors

The data per language was converted into scores that represent whether a language is embedded within a supportive context, ecosystem and climate giving it the possibility to flourish, or whether it may be without political will, funding, innovation and economic interest in the region. The score will, therefore, additionally indicate the probability of a language achieving DLE, given the assumption that a language in an environment with low support will also not be supported from a technical perspective any time soon.

The ELE core partners decided that in order for technological and contextual values to be compatible, a score between 0 and 1 would be assigned to the languages. With regard to the CFs specifically, 0 represents a context with hardly any potential for the development and collection of LTs and LRs, while 1 represents a high potential of those tasks being achieved.

To keep the metric as transparent as possible, it was decided to base the calculation on an average of the factors. Therefore, the intermediate goal was to calculate a score between 0 and 1 for each factor. The language with the lowest value for the respective factor is represented by a 0, and that with the highest value by a 1. The calculations performed to obtain those scores were the following:

1. Calculation of the range: highest value - lowest value = range;

[7] https://en.wikipedia.org/wiki/Languages_of_Ireland

2. $\frac{(value-minimum)*100}{range}$ = Percentage weighting of a language within the range;

3. The result is a relative value: To obtain a score between 0-1 the result is divided by 100;

4. Repetition of steps 1-3 for all languages and factors;

5. Calculation of the average of all factors per language.

The results achieved were weighted by three factors, namely the number of speakers, the scores based on the language status, and whether the language was an official language of the EU or not. The factors were considered to be highly relevant for the context to develop LTs and LRs due to the fact that the number of speakers has a big influence on the amount of investment by large companies, and the legal status or the EU status influences the amount of funding. The weighting included two steps: (i) the calculation of the average of the obtained overall scores, the scores for the number of speakers and the legal status, and (ii) the addition of 0.07 to the score for each official EU language. The second step was separated from the usual average calculation, because it would include two values: 0 for not being an official EU language and 1 for being an EU language. This results in a strong boost to every European language. Hence, English already had a score of around 0.7-0.8 without the boost, and smaller scores for the EU languages would have penalised English, which would not represent the reality of its dominant position today.

4.4. Selection of the Contextual Factors

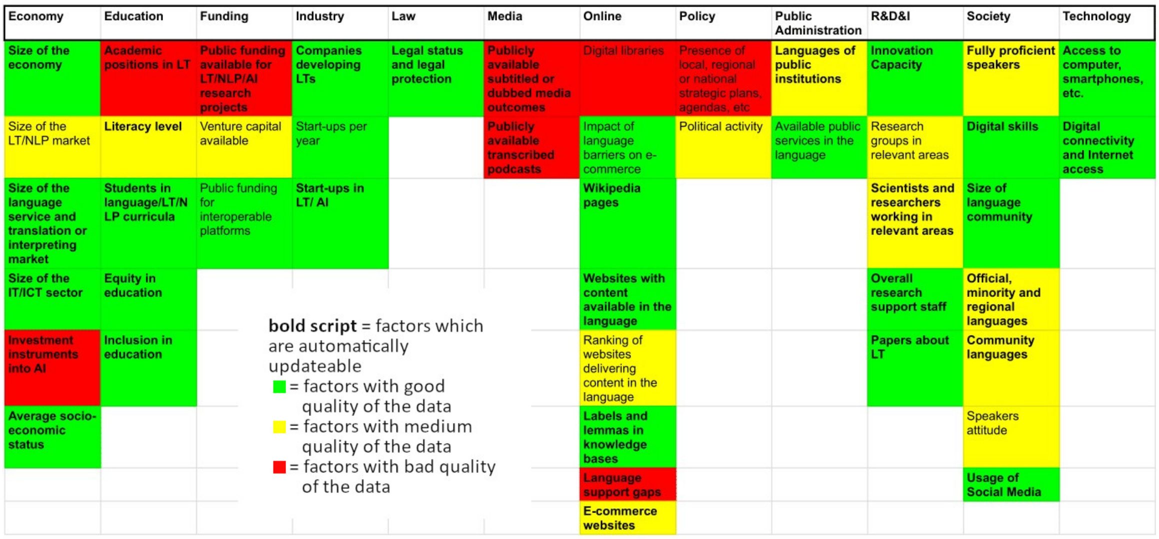
The described calculation approach allows the metric to be calculated with a different number of factors in order to either discover the best result or to compare different possible results. Accordingly, the factors were classified to select the contextual factors for the DLE metric based on two criteria: (i) the possibility of updating the data automatically, and (ii) the quality of the data. The quality of the data was chosen since biased data can highly distort the outcome of the metric, and the metric would show up any bias in the data. The possibility of updating the data automatically was selected as a criterion due to the fact that the metric will be implemented on the Dashboard of the ELG Catalogue, as explained in Section 2.3. In order to ensure that an up-to-date score for the languages is displayed, it is essential that the data can be automatically updated. In contrast, any data that has to be manually extracted will need a human expert to take care of its preparation after the end of the project. Given the fact that the project will end in June 2022, an automatically updating metric would be more convenient and future-proof its utility.

In Figure 2, the automatically updatable factors are highlighted in bold. This means that the data can be updated via an API of the source, or a script to gather structured information from websites. Any factors excluded via this criterion will need some manual preparation in order to be used. The criterion of the quality of the data separates the factors into three classes. They are highlighted through the different background colours in Figure 2. The factors highlighted in green are measured with data of good quality, those highlighted in yellow are based on medium data quality, and those in red having poor quality data. The yellow-marked factors were quantified with data with bigger gaps or with a missing variance in the data, which would give small differences a disproportionate impact on the score. The indicators of the red-highlighted factors were biased in a way, as some languages or countries had extreme outliers which were only explainable with distortion due, for example, to accuracy differences in the collection of the data.

Based on these criteria, the following configurations of contextual factors were examined:

1. All factors for which data was available: 46 factors

2. Those factors which are automatically updatable: 34 factors



Economy	Education	Funding	Industry	Law	Media	Online	Policy	Public Administration	R&D&I	Society	Technology
🟩 **Size of the economy**	🟥 **Academic positions in LT**	🟥 **Public funding available for LT/NLP/AI research projects**	🟩 **Companies developing LTs**	🟩 **Legal status and legal protection**	🟥 **Publicly available subtitled or dubbed media outcomes**	🟥 Digital libraries	🟥 Presence of local, regional or national strategic plans, agendas, etc	🟨 **Languages of public institutions**	🟩 **Innovation Capacity**	🟨 **Fully proficient speakers**	🟩 **Access to computer, smartphones, etc.**
🟨 Size of the LT/NLP market	🟨 **Literacy level**	🟨 Venture capital available	🟩 Start-ups per year		🟥 **Publicly available transcribed podcasts**	🟩 Impact of language barriers on e-commerce	🟨 Political activity	🟩 Available public services in the language	🟨 Research groups in relevant areas	🟩 **Digital skills**	🟩 **Digital connectivity and Internet access**
🟩 **Size of the language service and translation or interpreting market**	🟩 **Students in language/LT/NLP curricula**	🟩 Public funding for interoperable platforms	🟩 **Start-ups in LT/ AI**			🟩 **Wikipedia pages**			🟨 **Scientists and researchers working in relevant areas**	🟩 **Size of language community**	
🟩 **Size of the IT/ICT sector**	🟩 **Equity in education**					🟩 **Websites with content available in the language**			🟩 **Overall research support staff**	🟨 **Official, minority and regional languages**	
🟥 **Investment instruments into AI**	🟩 **Inclusion in education**					🟨 Ranking of websites delivering content in the language			🟩 **Papers about LT**	🟨 **Community languages**	
🟩 **Average socio-economic status**						🟩 **Labels and lemmas in knowledge bases**				🟨 Speakers attitude	
						🟥 **Language support gaps**				🟩 **Usage of Social Media**	
						🟨 **E-commerce websites**					

bold script = factors which are automatically updateable
🟩 = factors with good quality of the data
🟨 = factors with medium quality of the data
🟥 = factors with bad quality of the data

Figure 2: Classification of the Contextual Factors

3. Those factors with a good data quality: 26 factors
4. Those factors which are automatically updatable and with a good quality data: 21 factors
5. The factors were manually curated using four criteria: automatically updatable, having good quality data, not more than 2 factors per class, and a balance between the data types: 12 factors

The fewer factors included in the metric, the more likely it is that an important influencing factor will be omitted. However, the risk of distorting the metric by the data is reduced. The configurations with the data without good data quality showed such biases, in that individual languages obtained much better results than had been expected while other languages received very poor ones, even though the context of the language communities was considered to be more supportive. In the configuration with all factors, Emilian, Gallo, and Franco-Provencial, for example, achieved scores comparable to the official national languages without EU status or the lower-scored official EU languages. And the Basque language achieved one of the lowest scores, although the context for the regional language looks rather good compared to other regional languages. These results are strong evidence of bias.

4.5. Results

In all configurations that were examined, the top third is dominated by the official EU languages, while the regional and minority languages are presented as a long tail to the right. The official national languages which are not recognised as official EU languages are arranged between the official EU languages and the regional and minority languages. The results of configuration 5 with the 12 contextual factors can be viewed in Figure 3 and the results of configuration 3 with the factors with a good data quality in Figure 4. The results of configuration 4 lies in between the two presented results. The results of the configurations differ slightly in the score ratios between and within the three language groups. Note too that each coloured group features instances of single languages from adjoining groups: Serbian, in the green group, and Manx in the red group.

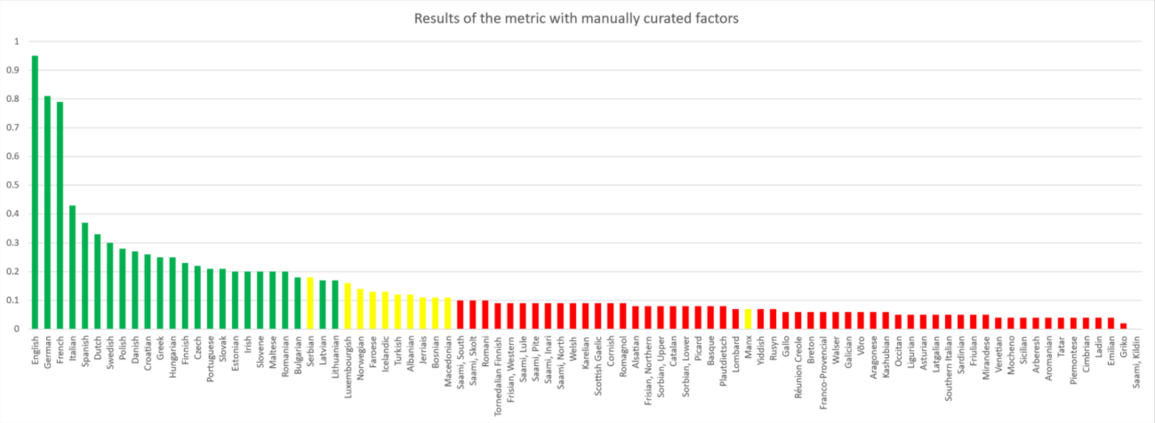


Figure 3: Results of the manually curated Contextual Factors

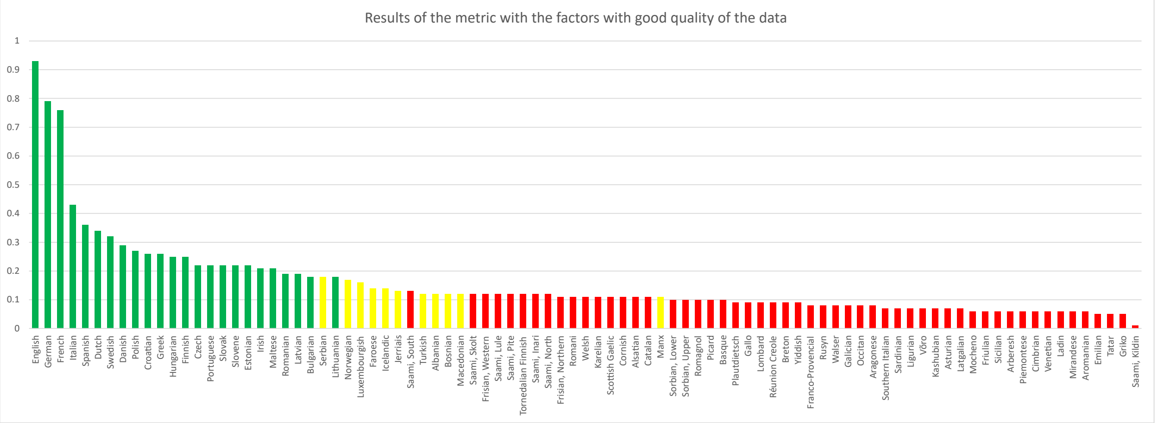


Figure 4: Results of the Contextual Factors with a good quality of the data

All configurations present English as having the best context for the development of LTs and LRs. English is followed by German and French, with German usually preceding French. Italian and Spanish are shown in positions 4 and 5. The position of Spanish with a worse score than Italian is caused by the inclusion of data from European countries only. If data had been included from countries outside of Europe, then Spanish, Portuguese, French and English would have had much higher scores given their prevalence in non-EU states. After the five leading languages, variations between the configurations begin to be seen. Mostly, Swedish, Dutch, Danish, Polish, Croatian, Hungarian and Greek are ranked in the upper half of the official EU languages. In some configurations, Finnish also joins this group. 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 score in keeping with the lower-scoring official EU languages, while Manx[8] is always presented as a downward outlier. Lan-

[8] Manx and Jerriais have been assigned to the group of national languages without being an official EU language, as both languages are recognised as official languages of Jersey and the Isle of Man. The classification of languages according to their official status is based on data from Ethnologue. Both islands are not part of the United Kingdom, although crown dependencies. Therefore, the two languages can be considered both official national languages or regional and minority languages.

guages such as Norwegian, Luxembourgish, Faroese and Icelandic achieve better scores than Albanian, Turkish, Macedonian and Bosnian. The scores for Jerriais are subject to comparatively large fluctuations, which is why the language is sometimes placed worse and sometimes better.

The regional and minority languages are usually led by Saami South and Skolt. Depending on the configuration, the languages Tornedalian Finnish, Romani, Northern and Western Frisian and the remaining Saami languages (apart from Saami, Kildin) achieve a score comparable to Saami, South and Skolt. 20 of the regional and minority languages achieve scores lower than 0.05 in configuration 5, while 31 of the languages obtain scores between 0.06 and 0.1. In the other configurations, the scores of the regional and minority languages are usually higher, but with similar differences between the scores of the languages. Saami, Kildin and Griko are the languages with the lowest scores.

4.6. Future Improvements

The results provided by the various CF configurations that were experimented with were sent to 37 partners from the project consortium by email, with a request for feedback. Furthermore, the results were presented at the General Assembly Meeting in early February to collect further input from the project partners to improve and finalise the results. A good amount of feedback was collected and evaluated as part of this consultation process. Languages mentioned multiple times as not being positioned where they should be according to the experts were Irish, Maltese, Croatian, Latvian, Norwegian, Icelandic, Farose, Jerriais and Manx. Moreover, the regional and minority languages Cornish, Scottish Gaelic, Emilian, Sicilian and most of the Saami languages were rated as not being placed in the correct relative position by at least one of the partners. Overall, the feedback considered 56 of the 89 languages.

In order to calculate a more suitable score for these languages, several suggestions were made. Since only pan-European data sources were taken into account, it is recommended to extend the data collection through national and regional sources. Additionally, it was pointed out that the context of languages spoken in countries outside of Europe is excluded in this analysis, and these missing statistics on the development of LTs and LRs would greatly impact the overall scores. Another suggestion refers to missing factors, such as the inclusion of the vitality status of the language being particularly important for regional and minority languages, or the integration of a factor representing the competition of national languages, if more than one official national language exist. Another idea is to replace the official EU status as a weighting factor with the country's membership in the European Economic Area (EEA), since countries within this alliance also have access to European research funds and networks.

Furthermore, ideas for preparing the data were submitted. These include, on the one hand, a stronger cleaning and standardisation of the data before they are processed and, on the other hand, the calculation of ratios between individual factors.

Suggestions were also made regarding the presentation of the results. Language communities having particularly complex political backgrounds are most likely to be biased by a simple calculation based on country data and should be highlighted and presented with the limits of data work for such cases. It is also suggested that languages functioning without a writing system are special cases for the development of LTs and LRs. This should be stressed in the presentation of the results.

Some of the feedback expressed reservations about the whole approach. Some partners pointed out that one methodology should not be used to take into account the different complex contexts given for the language communities. For example, languages like Maltese, Irish and the other Celtic languages, which scored better than expected according to our experts, are of note here. The relative prosperity of the United Kingdom seems to boost the regional and minority languages with the country-based data, although the reality of the language communities is strongly dominated by English. The same applies to Ireland, which has a strong economy, large ICT sector and significant investment in (English) AI and LT research & development, but a low level of support for Irish LT.

Another fundamental point of criticism is the inclusion of data not applied on a per capita basis. As a result, despite having relatively good support, some small language communities were unable to achieve a high score. The size of the language community has an impact on the economic interest, investment, number of researchers, etc. for the language, but for small language communities that have already invested a lot in their language and infrastructure, the scores obtained may appear too low.

These criticisms could certainly be debated at length, especially in the interest of finding effective solutions to the issues that were identified, but are difficult to avoid altogether with such a quantitative approach as the one that is required to define the CFs as part of the DLE Metric. A first stable result for the calculation of the CFs was achieved, which can be further refined and extended in future explorations.

5. Conclusions and Future Work

This deliverable has described the full operational specification of DLE and of the DLE Metric, based on a solid empirical basis that consists of the ELG Catalogue for the TFs and on a range of reputable high-quality sources for the CFs. This work is the result of the joint effort of the ELE Consortium and also takes on board the views of the broader LT and language-centric AI community in Europe, with which the ELE partners have extensive contacts. In the remainder of the ELE project, the DLE Metric will contribute to the formulation of the sustainable evidence-based SRIA and of the roadmap that will drive future efforts in equipping all European languages with the level of technological support that is needed to achieve full DLE in Europe by 2030, and will provide a transparent means to track and monitor progress in this direction. To this end, the deliverable has explained in detail the nature of the DLE Metric, emphasising in particular its dynamic and updatable nature, so that it can be used to track the progress of individual languages towards the ultimate goal of DLE for all European languages by 2030.

The deliverable has described the TFs and CFs that make up the DLE Metric; with regard to the TFs, the collaboration with the sister ELG project has been particularly valuable, in that the TFs rely on the data and accompanying metadata included in the ELG Catalogue as the ground truth and empirical foundation to measure and quantify the level of digital readiness of the languages covered by ELE. The overview of the TFs was accompanied by a discussion of the scoring and weighting mechanism adopted in the DLE metric, and a first possible implementation is illustrated to explain the overall design of the features and values that contribute to the TFs. As far as the CFs are concerned, the deliverable described the data sources that were identified and eventually chosen to extract the relevant indicators; part of this discussion was dedicated to the challenge of attributing proportionally the quantitative data on a per-country basis to the relevant languages and language communities. In addition, the deliverable reported on the data preparation process, foregrounding the importance of the quality of the data and of the possibility of refreshing the quantifiers underlying the CFs on a regular basis, to support the dynamic nature of the DLE Metric. We have also reported on a heuristic expert evaluation that was performed and provided useful indications to finalise the CFs, in addition to offering valuable suggestions for possible future improvements.

On the basis of this work that was conducted in close contact with the broader community in order to ensure that views from outside the consortium were also represented, the ELE Consortium is confident that the fully specified concept of DLE and its associated Metric proposed here represent valuable tools on which to base subsequent efforts to measure and improve the readiness of European languages for the digital age, also in the context of the formulation of the SRIA and roadmap. By drawing on the descriptive, diagnostic and predictive value of the DLE Metric, the community will have a solid and verifiable means of pursuing and evaluating much-needed developments in the interest of all European citizens. In this respect, we look forward to developments in the LT and language-centric AI community that will involve the DLE Metric, both during the lifespan of the ELE project as well as beyond. At the same time, we also hope that the DLE Metric will be recognised as a helpful tool by a range of key stakeholders at various local, regional, national and European levels who are committed to preventing the extinction of European languages under threat, and who are interested in promoting their prosperity in the coming years. Such stakeholders include decision- and policy-makers, industry leaders, researchers, and more generally citizens and societies across Europe. By virtue of being a tool to encourage and monitor the progress of all European languages to achieve the ambitious goal of full DLE by 2030, the DLE Metric serves purposes that can have a positive impact in several areas, including the economy as a whole with the Digital Single Market (DSM), industry, tourism, education and culture.

References

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. *arXiv*, 2021. 2110.06733.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon J. Greenhill, and Xia Hua. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 2021. doi: 10.1038/s41559-021-01604-y.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. Dataset Geography: Mapping Language Data to Language Users. *Computing Research Repository (CoRR)*, abs/2112.03497, 2021. Last accessed: 10.02.2021.

Federico Gaspari, Andy Way, Jane Dunne, Georg Rehm, Stelios Piperidis, and Maria Giagkou. D1.1 Digital Language Equality (preliminary definition). https://european-language-equality.eu/wp-content/uploads/2021/05/ELE_Deliverable_D1_1.pdf, 2021. Last accessed: 08.02.2022.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

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, 2020. European Language Resources Association (ELRA).

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. European Language Grid: An Overview. 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 3359–3373, Marseille, France, 2020. European Language Resources Association (ELRA).

Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault. The AI Index 2021 Annual Report. 2021. URL https://arxiv.org/abs/2103.06312.

Appendices

A. Full Specification of the Technological Factors of the DLE Metric

Table 1: First implementation of the Technological Factors of the DLE Metric – LRs

Features	Feature Values	Proposed Weights
Resource Type	corpus	5
	lexical conceptual resource	1.5
	language description	3.5
Subclass	raw corpus	0.1
	annotated corpus	2.5
	computational lexicon	2
	morphological lexicon	3
	terminological resource	3.5
	Wordnet	4
	Framenet	4
	model	5
	each of the others (there are 15 more)	0.5
Linguality Type	multilingual	5
	bilingual	2
	monolingual	1
Media Type	text	1
	image	3
	video	5
	audio	2.5
Annotation Type	*each of these - can be combined in a single LR*	0.25
Domain	*each of these - can be combined in a single LR*	0.3
Conditions of Use	other specific restrictions	0.5
	commercial uses not allowed	1
	no conditions	5
	derivatives not allowed	1.5
	redistribution not allowed	2
	research use allowed	3.5

Table 2: First implementation of the Technological Factors of the DLE Metric – Tools

Features	Feature Values	Proposed Weights
Language Independent	false	5
	true	1

Continued on next page

Table 2 – *Continued from previous page*

Features	Feature Values	Proposed Weights
Input Type	input text	2
	input audio	5
	input image	7.5
	input video	10
Output Type	output text	2
	output audio	5
	output video	10
	output image	7.5
	output numerical text	2.5
Function Type	text processing	3
	speech processing	10
	information extraction and information retrieval	7.5
	translation technologies	12
	human-computer interaction	15
	natural language generation	20
	support operation	1
	image/video processing	13
	other	1
	unspecified	1
Domain	*each of these – can be combined in a single tool*	0.5
Conditions of Use	unspecified	0
	other specific restrictions	0.5
	no conditions	5
	commercial uses not allowed	1
	derivatives not allowed	1.5
	research use allowed	3.5

B. Full Specification of the Contextual Factors of the DLE Metric

Table 3: List of Contextual Factors

Factors	Indicator(s)	Source
Size of the economy	Annual Gross Domestic Product (GDP)	Eurostat (2021). GDP and main components (output, expenditure and income). Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nama_10_gdp&lang=en
	GDP per capita	Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., and Hua, X. (2021). Global predictors of language endangerment and the future of linguistic diversity. Nature Ecology & Evolution, 6:163–173. https://doi.org/10.1038/s41559-021-01604-y
Size of the LT/NLP market	LT market in million Euro	European Commission, Directorate-General for Communications Networks, Content and Technology, Meertens, L., Choukri, K., Aguzzi, S., et al. (2019). Final study report on CEF automated translation value proposition in the context of the European LT market/ecosystem. https://data.europa.eu/doi/10.2759/142151
Size of the language service, translating or interpreting market	Number of organisations from the industry in the ELG catalogue	ELG Consortium (2021). Numbers about European Language Grid. Retrieved December 16, 2021, not published.
Size of the IT/ICT sector	Percentage of the ICT sector at the GDP	Eurostat (2021). Percentage of the ICT sector in GDP. Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_bde15ag&lang=en
	ICT service exports in Balance of Payment	World Bank (2021). ICT service exports By Country, in BoP, current US$ 1988-2019. Retrieved December 16, 2021, from https://wits.worldbank.org/CountryProfile/en/country/by-country/startyear/LTST/endyear/LTST/indicator/BX-GSR-CCIS-CD
Investment instruments into AI/ LT	Gross domestic expanditure on R&D in relevant areas[9]	Eurostat (2021). GERD by sector of performance and fields of R&D. Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=rd_e_gerdsc&lang=en

Continued on next page

[9] Relevant areas: Computer and information science, media and communications, humanities and language and

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Regional/ national LT market	No indicator found	
Average socio-economic status	Annual net earnings of a full-time single worker without children earning an average wage	Eurostat (2021). Annual net earnings of a full-time single worker without children earning an average wage. Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=earn_nt_netft&lang=en
	Life expectancy at age 60	Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., and Hua, X. (2021). Global predictors of language endangerment and the future of linguistic diversity. Nature Ecology & Evolution, 6:163–173. https://doi.org/10.1038/s41559-021-01604-y
Higher Education Institutions operating in the language	No indicator found	
Proportion of higher education conducted in the language	No indicator found	
Academic positions in relevant areas	Head count of the R&D personnel and researcher in relevant areas[10]	Eurostat (2021). R&D personnel and researchers by sector of performance, fields of R&D and sex. Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=rd_p_perssci&lang=en
Academic programmes of study in relevant areas	No indicator found	
Literacy Level	Literacy rate	SIL International (2021). Ethnologue. Retrieved December 16, 2021 from https://www.ethnologue.com/
Students in language/LT/NLP curricula	Total number of students in relevant areas[11]	Eurostat (2021). Distribution of graduates at education level and programme orientation by sex and field of education. Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=educ_uoe_grad03&lang=en

Continued on next page

literature.

[10] Relevant areas: linguistic and information and communication technology.

[11] Relevant areas: linguistic, information and communication technology.

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Equity in education	Tertiary educational attainment in percentage	Eurostat (2021). Population by educational attainment level, sex, age and degree of urbanisation (%). Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=edat_lfs_9913&lang=en
Inclusion in education	Percentage of foreigners who attain tertiary education	Eurostat (2021). Population by educational attainment level, sex, age and country of birth (%). Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=edat_lfs_9912&lang=en
Public funding available for LT/NLP/AI research projects	Number of projects funded by the European Commission in relevant areas[12]	European Research Council (2021). Statistics. Retrieved December 16, 2021, from https://erc.europa.eu/projects-figures/project-database
	Score from the National funding programs	LT-innovate (2016). National Funding Opportunities. Retrieved December 16, 2021, from https://www.lt-innovate.org/lt-observe/public-policy-observatory/national-funding-opportunities[13]
Venture capital available	Venture capital amounts in Euro	Bellucci, A., Gucciardi, G. and Nepelski, D. (2021). Venture Capital in Europe. Evidence-based insights about Venture Capitalists and venture capital-backed firms, EUR 30480 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-26939-7 (online), doi:10.2760/076298 (online), JRC122885.
Public funding for interoperable platforms	Number of platforms	Directorate-General for Research and Innovation (2020). Supporting the Transformative Impact of Research Infrastructures on European Research. Publications Office of the European Union, Luxembourg, ISBN 978-92-76-19271-8, doi: 10.2777/490221, KI-02-20-397-EN-C.

Continued on next page

[12] Relevant areas: Computer Science and Informatic, System and Communication Engineering, Cultures and Cultural Production.

[13] Source contains data on one web page per country.

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Companies developing LTs	Number of enterprises in the field of Information and Communication	Eurostat (2021). Annual enterprise statistics by size class for special aggregates of activities (NACE Rev. 2). Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=sbs_sc_sca_r2&lang=en
Start-ups per year	Percentage of "Enterprise births"	Directorate-General for Research and Innovation (2021). European Innovation Scoreboard 2021 Database. Retrieved December 16, 2021, from https://ec.europa.eu/docsroom/documents/46934
Start-ups in LT/ AI	Number of AI start ups	Degtyareva, G. (2017). Europe AI startups. Towards Data Science. Retrieved December 16, 2021, from https://docs.google.com/spreadsheets/d/1KaBg6qbGd4l66kahzz1pynW2Fv6ARJL55aAyMmmf9Ec/edit?usp=sharing
Copyright legislation and regulations	No indicator found	
Legal status and legal protection	Scores out of the legal status	SIL International (2021). Ethnologue. Retrieved December 16, 2021 from https://www.ethnologue.com/
Publicly available subtitled or dubbed visual media outcomes	Scores out of language transfer practices	Media-Commitee of LT-innovate (2008). Study on dubbing and subtitling needs and practices in the European audiovisual industry. Retrieved December 16, 2021, from http://www.lt-innovate.org/lt-observe/document/study-dubbing-and-subtitling-needs-and-practices-european-audiovisual-industry
	Scores out of the answers about broadcast practizes	European Federation of National Institutions for Language (2019) European Language Monitor 4. Retrieved December 16, 2021, from https://juniper.nytud.hu/elm4/index
Publicly available transcribed podcasts	Number of entries in the digital library of cba	Cultural broadcasting archive (2021). Suche. Retrieved December 16, 2021, from https://cba.media/explore
Digital libraries	Percentage of contribution to Europeana	European Commission (2009). EUROPEANA – Europe's Digital Library: Frequently Asked Questions. MEMO/09/366, Brussels. Retrieved December 16, 2021, from https://ec.europa.eu/commission/presscorner/detail/en/MEMO_09_366

Continued on next page

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Impact of language barriers on e-commerce	Percentage of population buying cross-border	STOA (2017). Language equality in the digital age – Towards a Human Language Project. Retrieved December 16, 2021, from http://www.europarl.europa.eu/stoa/
Digital literacy	No indicator found	
Wikipedia pages	Number of articles in Wikipedia	Wikimedia (2021). List of Wikipedias. Retrieved December 16, 2021, from https://meta.wikimedia.org/wiki/List_of_Wikipedias
Websites with content available exclusively in the language	No indicator found	
Websites with content available in the language (but not exclusively)	Percentage of websites in the languages	W3Tech (2021). Usage statistics of content languages for websites. Retrieved December 16, 2021, from https://w3techs.com/technologies/overview/content_language
Web pages	No indicator	
Ranking of websites delivering content	Matrix of the 12 selected websites supporting the languages	STOA (2017). Language equality in the digital age – Towards a Human Language Project. Retrieved December 16, 2021, from http://www.europarl.europa.eu/stoa/
Labels and lemmas in knowledge bases	Number of lexemes in Wikipedia	Wikimedia (2021). Wikidata:Lexicographical data/Statistics/Counts of various things by language. Retrieved December 16, 2021, from https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Statistics/Counts_of_various_things_by_language
Language support gaps	Language matrix of supported features	W3C (2017). Language matrix: International typography on the Web. Retrieved December 16, 2021, from https://www.w3.org/International/typography/gap-analysis/language-matrix.html
E-commerce websites	T-Index	Imminent (2021). Austria. Retrieved December 16, 2021, from https://imminent.translated.com/data-index/austria[14]

Continued on next page

[14] Source contains data on one web page per country.

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Presence of local, regional or national strategic plans, agendas, etc.	Scores out of a list of the published national AI strategies	European Commission, Joint Research Centre, Organisation for Economic Co-operation and Development, Van Roy, V., Rossetti, F., Perset, K., et al. (2021). AI watch, national strategies on artificial intelligence : a European perspective, Publications Office, https://data.europa.eu/doi/10.2760/069178
	Scores out of the question: Is there, in your country, an official language plan/strategy published by your country's government or some organisation close to the government?	European Federation of National Institutions for Language (2019) European Language Monitor 4. Retrieved December 16, 2021, from https://juniper.nytud.hu/elm4/index
Recognition and promotion of the LR ecosystem	No indicator found	
Consideration of regional or national bodies for the citation of LRs	No indicator found	
Promotion of regional, national or international cooperation	No indicator found	
Public and community support for the definition and dissemination of resource production best practices	No indicator found	
Policies to provide, maintain and update BLARKs	No indicator found	
Political activity	Scores out of the list of documents, initiatives, etc. regarding LTs	Aldabe, I., Rehm, G., Rigau, G. and Way, A. (2021). D3.1 Report on existing strategic documents and projects in LT/AI. Retrieved December 16, 2021, from https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf

Continued on next page

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Languages of public institutions	Number how many constitutions are written in the language	European Federation of National Institutions for Language (2019) European Language Monitor 4. Retrieved December 16, 2021, from https://juniper.nytud.hu/elm4/index
Available public services in the language	Percentage of a maximum score about digital public services	European Commission (2021). Digital Economy and Society Index (DESI) 2021: Thematic chapters. Retrieved December 16, 2021, from https://digital-strategy.ec.europa.eu/en/policies/desi
	Score for digital public services for citizens	European Commission (2021). Digital Economy and Society Index (DESI) 2021: Thematic chapters. Retrieved December 16, 2021, from https://digital-strategy.ec.europa.eu/en/policies/desi
Innovation capacity	Innovation Index	Directorate-General for Research and Innovation (2021). European Innovation Scoreboard 2021 Database. Retrieved December 16, 2021, from https://ec.europa.eu/docsroom/documents/46934
Research groups in LT	Number of research organisations	ELG Consortium (2021). Numbers about organisations per country. Retrieved December 16, 2021, not published.
Research groups/ companies predominantly working on the respective language	No indicator found	
Research & Development staff involved in LT	No indicator found	
Suitably trained and qualified Research & Development staff in LT	No indicator found	
Capacity for talent retention in LT	No indicator found	
State of play of NLP/AI at large	No indicator found	
Scientists and researchers working in LT/ on the language	Total number of researchers in relevant areas[15]	Eurostat (2021). R&D personnel and researchers by sector of performance, fields of R&D and sex. Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=rd_p_perssci&lang=en

Continued on next page

[15] Relevant areas: Computer and information science, media and communication, humanities, languages and lit-

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Researchers and scholars whose work benefits from the availability of LRs and LTs	No indicator found	
Overall research support staff	Dead count of the research support staff	Eurostat (2021). R&D personnel by sector of performance, professional position and sex. Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=rd_p_persocc&lang=en
Scientific associations or general scientific and technology ecosystem	No indicator found	
Papers about LT and or the language	Number of papers about Machine Translation, Speech Synthesis and Information Retrieval	European Commission, Directorate-General for Communications Networks, Content and Technology, Meertens, L., Choukri, K., Aguzzi, S., et al. (2019). Final study report on CEF automated translation value proposition in the context of the European LT market/ecosystem, https://data.europa.eu/doi/10.2759/142151
Number of papers reporting studies on language	Number of references	Hammarström, H., Forkel, R., Haspelmath, M. and Bank, S. (2021). Glottolog 4.5. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5772642 Retrieved December 16, 2021, from https://glottolog.org/
Importance, relevance or recognition of the language	No indicator found	
Fully proficient (literate) speakers	Number of L1 speakers	SIL International (2021). Ethnologue. Retrieved December 16, 2021 from https://www.ethnologue.com/
Digital Skills	Percentage of individuals with basic or above basic overall digital skills	Eurostat (2021). Individuals' level of digital skills (until 2019). Retrieved December 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_sk_dskl_i&lang=en
Size of language community	Total number of speakers	SIL International (2021). Ethnologue. Retrieved December 16, 2021 from https://www.ethnologue.com/[16]

Continued on next page

erature.

[16] Gaps filled and some numbers corrected as explained in Section 4.2.

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Population that does not speak the official language(s)	No indicator found	
Official languages and recognized minority and regional languages	Total number of the languages with an official status	SIL International (2021). Ethnologue. Retrieved December 16, 2021 from https://www.ethnologue.com/
	Number of bordering languages	Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., and Hua, X. (2021). Global predictors of language endangerment and the future of linguistic diversity. Nature Ecology & Evolution, 6:163–173. https://doi.org/10.1038/s41559-021-01604-y
Community languages	Number of Community languages	SIL International (2021). Ethnologue. Retrieved December 16, 2021 from https://www.ethnologue.com/
Available time resources of the members of the language community	No indicator found	
Civil society stakeholders working on (preserving) the respective language	No indicator found	
Speakers' attitudes towards the language	Total number of participants wanting to acquire the language	Directorate-General for Communication (2014). Special Eurobarometer 386: Europeans and their Languages. Retrieved December 16, 2021, from https://data.europa.eu/data/datasets/s1049_77_1_ebs386?locale=en
Involvement of indigenous peoples	No indicator found	
Sensitivity to barriers that impede the availability of new technology, content and services	No indicator found	
Usage of Social Media or networks	Total number of social media users	Kepios (2021). Datareportal. Retrieved December 16, 2021, from https://datareportal.com/reports?offset=1613118017367&tag=Digital+2021
	Percentage of social media users	Kepios (2021). Datareportal. Retrieved December 16, 2021, from https://datareportal.com/reports?offset=1613118017367&tag=Digital+2021

Continued on next page

Table 3 – *Continued from previous page*

Factors	Indicator(s)	Source
Open-source technologies of LTs	No indicator found	
Access to computer, smartphone etc.	Percentage of households with access to a computer from home	OECD (2021). Access to computers from home (indicator). doi: 10.1787/a70b8a9f-en, retrieved December 16, 2021.
Digital connectivity and Internet access	Percentage of households with broadband access	Eurostat (2021). Households with broadband access. Retrieved 16, 2021, from https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_r_broad_h&lang=en

Table 4: Conversion from plain text into scores

Factor	Merging of the scores	Conversion from text to scores
Public funding available for LTs	Adding up of the scores for each country	1 for regional funding 1 for national funding 1 for intranational funding 1 for ESIF 1 for EUREKA 1 for EUROSTAT
Legal status and legal protection	Adding up of the scores per language	10 for statutory national language 10 for de facto national working language 2 for statutory provincial language 2 for statutory provincial working language 1 for recognized language
Publicly available media outcomes	Adding up of two scores: one score for language transfer practices for cinema works screened and one for television works broadcast	2 for dub 1.5 for voice over 1.5 for sub and dub 1 for sub

Continued on next page

Table 4 – *Continued from previous page*

Factor	Merging of the scores	Conversion from text to scores
	Adding up of the scores + division through the number of answers	Broadcast in original language: 5 for mostly/ always, 2.5 for sometimes Broadcast with dubbing: 4 for mostly/ always, 2 for sometimes Broadcast in original language with voice-over: 3 for mostly/ always, 1.5 for sometimes Dual-channel sound: 2 for mostly/ always, 1 for sometimes Broadcast with subtitles: 1 for mostly/ always, 0.5 for sometimes
Presence of local, regional or national strategic plans	One of the score per country	1 for no plan/ strategy 2 for a plan without mentioning LT 3 for a plan mentioning LT 4 for a plan mentioning LT and minority and regional languages
Political activity	Adding up of the scores per country	1 score for each document 1 score for each document mentioning LT 2 for each document exclusively about LT 1 for a document covering a specific language 2 for each document published 2020/2021 1 for each document published 2019/2018