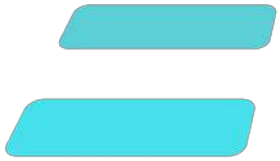

EUROPEAN LANGUAGE EQUALITY

D1.1

Digital Language Equality (preliminary definition)

Authors	Federico Gaspari (DCU), Andy Way (DCU), Jane Dunne (DCU), Georg Rehm (DFKI), Stelios Piperidis (ILSP), Maria Giagkou (ILSP)

Dissemination level	Public
Date	31-03-2021

About this document

Project	European Language Equality (ELE)
Grant agreement no.	*to be decided*
Coordinator	Prof. Andy Way (DCU)
Co-coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.1
Deliverable title	Digital Language Equality (preliminary definition)
Type	Report
Number of pages	25
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 31-03-2021– Actual: 31-03-2021
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.1 Defining Digital Language Equality
Authors	Federico Gaspari (DCU), Andy Way (DCU), Jane Dunne (DCU), Georg Rehm (DFKI), Stelios Piperidis (ILSP), Maria Giagkou (ILSP)
Reviewers	German Rigau (UPV/EHU), Kepa Sarasola (UPV/EHU)
EC project officer	Aleksandra Wesolowska
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	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

1. Introduction **1**
1.1. Background and Context . . . 1
1.2. Main Aims . . . 1
1.3. Consortium . . . 2

2. Preliminary Definition of Digital Language Equality **2**
2.1. Key Features . . . 3
2.2. Technological Factors . . . 4
2.2.1. Tools and Services . . . 5
2.2.2. Corpora . . . 5
2.2.3. Language Descriptions and Models . . . 6
2.2.4. Lexical and Conceptual Resources . . . 6
2.2.5. Projects . . . 6
2.2.6. Organizations . . . 7
2.3. Six Indicators of Linguistic Digital Readiness . . . 7
2.4. Contextual Factors . . . 7

3. Open Issues and Challenges **9**
3.1. Size of Language Resources . . . 9
3.2. Year and Contemporaneity of Language Resources . . . 9
3.3. Maintenance and Updates . . . 10
3.4. Cost, Licensing and Access . . . 10

4. Conclusions and Next Steps **11**

A. Technological Factors **13**

B. Contextual Factors **15**

List of Figures

1. Intersection between technological and contextual factors 8
2. Computing the Digital Language Equality metric 11

List of Tables

1. Digital Language Equality – Technological factors 13
2. Digital Language Equality – Contextual factors 15

List of Acronyms

AI	Artificial Intelligence
BLARK	Basic Language Resource Kit
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CS	Computer Science
CULT	European Parliament's Committee on Culture and Education
DL	Deep Learning
DLE	Digital Language Equality
ECSPM	European Civil Society Platform for Multilingualism
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELG	European Language Grid (EU project, 2019-2022)
ELM	European Language Monitor
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
EP	European Parliament
GA	Grant Agreement
HLT	Human Language Technology
ICT	Information and Communication Technology
IT	Information Technology
ITRE	European Parliament's Committee on Industry, Research and Energy
LDR	Linguistic Digital Readiness
LR	Language Resource/Resources
LRT	Language Resource/Resources and Technology/Technologies
LSP	Language Service Provider
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NCC	National Competence Centre
NLP	Natural Language Processing
NLU	Natural Language Understanding

SRA	Strategic Research Agenda
SRIA	Strategic Research and Innovation Agenda
SSH	Social Sciences and the Humanities
STOA	Science and Technology Options Assessment
TRL	Technology Readiness Level
WP	Work Package

Abstract

This deliverable provides a preliminary transparent working definition of digital language equality (DLE) based on a set of modular quantifiers, measures or indicators, to accurately reflect the level of support of language technologies for European languages as an essential requirement of DLE for the present as well as for the future. The deliverable is structured as follows. Section 1 presents the background and context of the ELE project, clarifies its main aims, and provides a brief overview of the consortium. Section 2 introduces the preliminary working definition of Digital Language Equality and the DLE metric, first explaining the rationale behind these concepts and then presenting their key features. This part includes a discussion of the technological and contextual factors on which the preliminary definition is based, which are also used to compute the DLE metric; the six technological factors make up a set of linguistic digital readiness indicators, and can be compared across languages at a fine-grained level. Section 3 reviews some of the main open issues and challenges to be addressed in subsequent stages of the project with a discussion of the approach taken with regard to the size of Language Resources, their year of release and the issues of contemporaneity, versions and updates, as well as on the aspects of costs, licensing and access. Finally, Section 4 briefly draws some conclusions and explains how this Deliverable D1.1 will be expanded upon in the subsequent Deliverable D1.3.

1. Introduction

1.1. Background and Context

In a plenary meeting on 11 September 2018, the European Parliament adopted by an overwhelming majority a joint ITRE/CULT report, *Language equality in the digital age*, with a resolution that included over 40 recommendations. These concerned the improvement of the institutional framework for language technology policies at EU level, EU research policies, education policies to improve the future of Language Technologies (LTs) in Europe, and the extension of the benefits of LTs for both private companies and public bodies (European Parliament, 2018). In particular, the resolution highlighted many important areas that are of direct interest to the ELE project, e.g., it called on the Commission "to establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands". Against this background, the ELE project addresses some of these recommendations and lays the foundations to elicit an evidence-based roadmap with strong support from the wider community that can provide the basis for a systematic plan and strategic agenda to achieve full DLE in Europe by 2030.

1.2. Main Aims

While the 24 official EU languages are granted equal status politically, technologically they are far from equally supported; in addition, there are several regional and minority languages that have traditionally suffered from limited support, especially to future-proof their use and very existence in the digital age. *The goal of ELE is the systematic and inclusive development of an all-encompassing strategic research, innovation and implementation agenda and roadmap for achieving full digital language equality in Europe by 2030.* To reach this ambitious objective, the project will draw up a sustainable evidence-based strategic research agenda and roadmap setting out actions, processes, tools and actors to achieve full DLE of all languages used in Europe through the effective use of LTs.

A fundamental element of this effort is the definition of DLE, whose preliminary formulation is presented in Section 2.1, which underpins the activities of other work packages (WPs) and tasks for the entire project. The preparation of the plan to achieve DLE in Europe by 2030 requires the accurate and up-to-date description of the current state of technology support for Europe's languages and the identification of gaps and issues with regard to LTs, also considering neighbouring disciplines.

1.3. Consortium

The ELE consortium is ideally positioned to pursue the ambitious goals of the project. It consists of a total of 52 members: five core partners, nine networks, associations and initiatives, nine companies and 29 research organisations. In addition to all official European languages, the partners' expertise covers several unofficial, regional and minority languages, either through consortium partners or through the umbrella organisations ELEN and EC-SPM. The consortium as a whole brings together research and industry partners as well as wider networks representing a very broad range of stakeholders that have come together to achieve full DLE for all European languages by 2030.

2. Preliminary Definition of Digital Language Equality

This report presents the preliminary conceptualization of DLE, through an initial transparent working definition based on a set of modular quantifiers, measures and indicators. This definition underpins the work of the entire project. The fundamental question that has guided its formulation is: "What LTs are required to achieve DLE in Europe by 2030?". While the proposed definition is firmly rooted in the state-of-the-art, it will also serve the needs of the languages targeted in the project and the expectations of the relevant language communities in the future. The preliminary definition provides the basis for the full specification of DLE that will be presented in Deliverable D1.3, to be submitted in January 2022.

Language "equality" does not mean "sameness" on all counts, regardless of the respective environments; we recognize the different historical developments and current situations of the very diverse languages that are targeted in the project, along with their specific features, different needs and realities of their communities, e. g., in terms of number of speakers, range of uses of the languages, etc., which inevitably vary significantly. It would be naïve and unrealistic in practice to ignore these facts, and to set out to erase the differences that make languages truly unique, as key components of the heritage and as a vital reflection of the communities that use them. This is also a core element of multilingualism in Europe, where all languages are valued as inherent components of the social fabric that connects European citizens in their diversity. The challenge tackled head on by ELE is to enable all languages, regardless of their specific circumstances, to realize their full potential, supporting them in achieving full digital equality in the coming decade.

The notion of DLE does not involve any judgement of the political, social and cultural status or value of the languages, insofar as they all collectively contribute to a multilingual Europe, that should be supported and promoted. Alongside the fundamental concept of equality, we also recognize the importance of the notion of equity, meaning that for some languages, and for some needs, a specific effort is necessary. Specific access to certain services and resources (for example to revitalize a language, or to promote the development of education through that language) is very important for some of Europe's languages.

The preliminary definition of DLE is articulated in Section 2.1, on the basis of an explanation of its key features and of how it will be operationalized in the project. After presenting the technological factors (Section 2.2) that make up six specific indicators of linguistic digital

readiness (LDR) that can be compared across languages (Section 2.3), the discussion presents the contextual factors contributing to DLE, that reflect the situation in which the languages are used in terms of economy, education, society, etc. (Section 2.4).

2.1. Key Features

Below we present the key features of the preliminary definition of DLE. First of all, the definition was designed to be modular and flexible, i. e., consisting of well-defined (separate and independent, but tightly integrated) quantifiers, measures and indicators, selected to ensure compatibility and interoperability with the metadata schema of the European Language Grid (ELG)[1] (Labropoulou et al., 2020; Rehm et al., 2020) which plays a crucial technical role in the project. ELG develops a cloud platform that bundles together functional software, data sets, corpora, repositories and applications to benefit European society, industry and academia and administration, while also addressing the fragmentation of the European LT landscape by providing a convenient single access point to the European LT community and its offerings. The DLE definition adopted in ELE is fully aligned with ELG. The report on language equality in the digital age on which the resolution adopted by the European Parliament mentioned in Section 1.1 (European Parliament, 2018) was based stated that LTs have not been adequately considered in the past, despite various investments, and they should be given due attention in the future; the ELE project aims to contribute to this vision.

The definition of DLE drew inspiration from the Basic Language Resource Kit (BLARK),[2] which has been used to define what LRs are needed, and to classify what already exists and what may need to be produced for specific languages. Having such information allows for a more realistic estimation of costs and efforts required for future LR production for specific languages (Krauwer, 2003). BLARK has been a useful instrument in the past for human language technology (HLT) experts to help report on the coverage of a certain language and thus plan for future actions. Due to the modular and flexible design of the preliminary DLE definition, the chosen quantifiers, measures and indicators can be dynamically adapted and updated in the future, leading to the final specification of the definition. The definition also supports the remapping of the contributing factors and criteria based on updated (more fine-grained or, conversely, coarser) or revised schemata of the metadata adopted in ELG. The preliminary definition can also accommodate additional elements to be introduced on the basis of evidence and feedback gathered during the project (e. g., as part of the landscaping exercises and technical deep-dives, as well as based on input from the wider community).

We have tried to make the definition transparent and similarly intuitive for linguists, LT experts, activists, policy-makers and European citizens at large, to encourage its widest possible uptake and buy-in from the broader community. While we wanted the definition to be founded on solid, widely agreed, principles, we also aimed at striking a balance between a methodologically sound and theoretically convincing definition, and a formulation that would also be applicable for computing the DLE metric, as well as usable in the community, including to inform future language policies – and LT policies – at the local, regional, national and European levels. The definition can be used to guide and prioritize future efforts in LT development and LR creation, collection, and curation activities. Through data modelling, analytics and visualization, languages facing similar challenges to achieve full digital equality can be grouped together, and requirements can be formulated to support them in remedying the existing gaps and advancing towards full DLE in the coming decade.

[1] https://live.european-language-grid.eu
[2] http://www.blark.org

Digital Language Equality – preliminary definition

Digital Language Equality (DLE) is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age.

The definition of DLE will be used to compute an easy-to-interpret metric for individual languages. The DLE metric will enable the quantification of the level of technological support for each language in scope of the project and, crucially, the straightforward identification and visualization of the gaps and shortcomings that hamper the achievement of full DLE. This approach enables direct comparisons across languages, tracking their advancement towards the goal of DLE, as well as the prioritization of needs, especially to fill existing gaps, focusing on realistic and feasible targets.

Digital Language Equality (DLE) Metric – preliminary definition

The **Digital Language Equality (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 DLE Metric is computed for each language on the basis of various factors, grouped into technological support (technological factors, e. g., the available language resources, tools and technologies) and a range of situational context factors (e. g., societal, economic, educational, industrial).

In Europe, it facilitates the identification and prioritization of needs, the preparation and planning of strategic activities and policies as well as the formulation of research, development and innovation objectives, with the overall goal of promoting the achievement of full DLE in Europe by 2030. The DLE metric will capture the needs and expectations of the various European languages and the shortfalls with respect to being adequately supported in the digital age so as to achieve equality. Later on in the process, specific scores, relative weights and perhaps also penalties will be assigned to the various factors involved in the computation of the DLE metric: this will result in up-to-date language-specific indicators that are comparable with one another, which will allow us to track the progress of technology support for the various languages, and all respective indicators, towards the goal of DLE in Europe by 2030. By virtue of being applicable across languages, the metric score provides six specific indicators of LDR, which are explained in more detail in Sections 2.2 and 2.3, showing their relative degree of maturity in terms of technological support and achievement of digital equality.

2.2. Technological Factors

On the basis of the preliminary definition of DLE (Section 2.1), the first set of quantifiers to consider are technological factors, that concern the availability of Language Resources and Technologies (LRTs), as well as the projects and organizations covering specific languages. Following the ELG categorization and metadata schema, these technological factors are divided into six main categories, namely (i) tools and services, (ii) corpora, (iii) language models and computational grammars (collectively referred to as language descriptions), (iv) lexical and conceptual resources, (v) projects, and (vi) organizations. They are listed in full in Appendix A and described in the following sections.

2.2.1. Tools and Services

This first category of technological factors is defined as tools and services offered via the web or running in the cloud, but also downloadable tools, source code, etc. They include for example basic NLP tools for the European languages (morphological analysers, part-of-speech taggers, lemmatizers, parsers, etc.); authoring tools (e. g. spelling, grammar and style checkers); services for information retrieval, extraction, and mining, text and speech analytics, machine translation, natural language understanding and generation, speech technologies, conversational systems, etc. For the purposes of computing the DLE metric, these are the relevant quantifiers that apply to tools and services, whose scoring and weighting mechanism will be defined in D1.3:

- Language(s)
- Domain(s)
- Licence
- Type of access
- Function(s) / Task(s)
- Language dependent
- Language(s) of input/output
- Media type(s) of input/output

A relevant quantifier to take into consideration for the DLE metric with regard to tools and services is the technology readiness level (TRL).

2.2.2. Corpora

This second category is defined as corpora or datasets, including collections of text documents, text segments, audio transcripts, audio and video recordings, etc., monolingual or bi-/multilingual, raw or annotated. For the purposes of computing the DLE metric, these are the relevant quantifiers that apply to corpora, whose scoring and weighting mechanism will be defined in D1.3:

- Language(s)
- Domain(s)
- Licence
- Type of access
- Corpus subclass
- Media type(s) of parts
- Multilinguality type
- Corpus size (based on corpus size unit)

2.2.3. Language Descriptions and Models

The third category is described as language models and computational grammars. For the purposes of computing the DLE metric, these are the relevant quantifiers that apply to language descriptions and models, whose scoring and weighting mechanism will be defined in D3.1:

- Language(s)
- Domain(s)
- Licence
- Subclass of grammar/model

2.2.4. Lexical and Conceptual Resources

This fourth category (i. e., lexical and conceptual resources) includes resources organised on the basis of lexical or conceptual entries (lexical items, terms, concepts etc.) with their supplementary information (e. g., grammatical, semantic, statistical information, etc.). They comprise computational lexica, gazetteers, ontologies, term lists, thesauri, etc. For the purposes of computing the DLE metric, these are the relevant quantifiers that apply to lexical and conceptual resources, whose scoring and weighting mechanism will be defined in D1.3:

- Language(s)
- Domain(s)
- Licence
- Lexical/conceptual resource subclass
- Media type(s) of parts
- Encoding level
- Number of entries (size)

2.2.5. Projects

The fifth category consists of projects that have funded or developed LRTs. For the purposes of computing the DLE metric, these are the relevant quantifiers that apply to funded projects, whose scoring and weighting mechanism will be defined in D1.3:

- Language(s)
- Domain(s)
- Number of consortium partners
- Technology sectors/areas/specialisms
- Duration
- Budget

2.2.6. Organizations

The sixth, and final, category of technological factors concerns organizations that are or have been active in the LT community and landscape in Europe, especially with regard to developing LTs or LRs or conducting research in the wider area of Language Technology. For the purposes of computing the DLE metric, these are the relevant quantifiers that apply to organizations, whose scoring and weighting mechanism will be defined in D3.1:

- Language(s)
- Domain(s)
- Type
- Technology sectors/areas/specialties
- Number of people working in the organisation
- Number of individual members
- Number of corporate/institutional members

2.3. Six Indicators of Linguistic Digital Readiness

The overall score of the DLE metric will be computed on the basis of the weighted scores assigned to the technological factors. The six categories of technological factors also serve as distinct fine-grained indicators of linguistic digital readiness (LDR) that can be specifically compared (like with like) across languages.

2.4. Contextual Factors

The second set of measures contributing to the DLE metric consists of contextual factors, that do not refer to strictly technological, linguistic or language-related indicators, but rather have to do with general conditions and situations of the broader context. These have been inspired by a number of diverse sources and past projects, most notably: the STOA (2017) report, which promoted goals that are well aligned with those of ELE, and also proposed and assessed a set of institutional, research, industry, market and public service policy options; the META-NET White Paper series *Europe's Languages in the Digital Age* (Rehm and Uszkoreit, 2012);[3] EFNIL's European Language Monitor (ELM), e. g., the most recent ELM4 from 2019 includes some relevant features;[4] both the FLaReNet report (Calzolari et al., 2011) and the META-NET Strategic Agenda for Multilingual Europe 2020 (Rehm and Uszkoreit, 2013) discussed some of the key elements needed to achieve language equality that included some useful starting points; the project on Modeling the Linguistic Consequences of Digital Language Contact;[5] finally, the Digital Language Diversity Project[6] worked on the identification of indicators and parameters for the computation of the level of technological readiness of languages, which eventually resulted in a set of indicators and their mapping on a scale of digital language vitality.[7]

In addition to these main sources that were consulted alongside similar relevant projects and initiatives, the list of contextual factors was also formulated on the basis of a consultation

[3] http://www.meta-net.eu/whitepapers/
[4] http://efnil.org/projects/elm
[5] http://molicodilaco.hi.is
[6] http://www.dldp.eu
[7] http://wp.dldp.eu/digital-language-vitality-scale/

of all ELE partners, in the interest of a comprehensive and inclusive range of contextual factors. The factors that were eventually selected refer to the region/country/countries/area(s) of the specific language in question, whatever level of granularity and specificity is most appropriate at the geographic and/or geopolitical level in connection with the relevant language. Clearly, depending on the spread and reach of the languages that are considered, very different considerations may apply to their links with one or more particular regions, countries or (supranational) areas. Appendix B lists the 72 contextual factors identified so far as a result of this consultation process, that have been clustered into 12 categories, namely:

- Economy
- Education
- Funding
- Law
- Media
- Policy
- Public administration
- Society
- Industry
- Online
- Research & Development & Innovation
- Technology

As shown in Figure 1, some of the technological factors (in particular projects and organizations) and some of the contextual factors (i. e., industry, online, research & development & innovation as well as technology) fall somewhat between and occupy some common ground. It is envisaged that subsequent activities in the project will help refine the list of contextual factors, assigning an appropriate scoring and weighting mechanism to them to compute the DLE metric. Their respective scoring systems and weights will be thoroughly described in the full final specification of DLE in Deliverable D1.3.

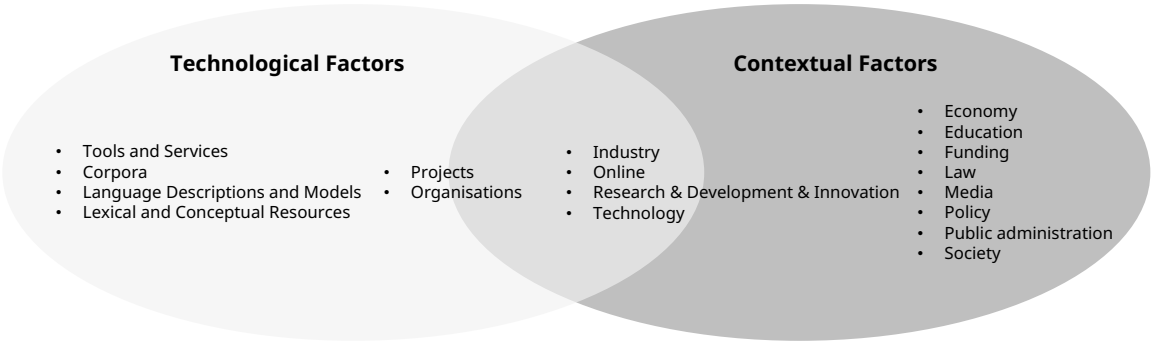

Figure 1: Intersection between technological and contextual factors

Taking a broader and longer-term view, beyond the ELE project, the data points related to these contextual measures will be crucial for defining the next steps, e. g. to foster collaboration between research centres and groups, including internationally and across regions. The consideration of the technological factors and contextual factors will represent key elements to design an effective plan to foster EU-wide collaboration in the LT space to fully realize the vision of complete DLE for all European languages by 2030.

3. Open Issues and Challenges

On the basis of the preliminary definition of DLE (Section 2.1), this section reviews some of the open issues and challenges to be addressed going towards the final full specification of DLE, including the exact scores, penalties and weights to be applied. Detailed guidelines on how to address the open issues and challenges included below are elaborated in the *Guidelines for Task 1.3 Contributors of ELE* (Giagkou and Piperidis, 2021).

3.1. Size of Language Resources

In general, there are benefits to setting a minimum size criterion to make language resources such as corpora or grammars acceptable as part of the collection effort undertaken in ELE, e. g., to avoid collecting small resources that cannot be realistically used in concrete technology development scenarios. However, it would be difficult to establish arbitrarily what this minimum size threshold should be, also considering the variety of technological factors to be considered (Section 2.2), and the specifics of the several languages to be compared with each other in terms of LDR. In the end, the pragmatic decision was made not to set any minimum size requirement. The thinking behind this choice was that relatively small data sets are common in less-resourced languages, for particular domains, etc., while, for example, there is the possibility to merge small but homogeneous data sets to create bigger ones that would, in fact, be useful, e.g., in domain adaptation for machine translation. This approach is in line with other European projects and initiatives, such as ELRI[8] that is now over, ELRC[9], CURLICAT[10] and PRINCIPLE,[11] which are currently ongoing, and similar efforts at national and regional level across Europe. Consistently with the ethos of these related initiatives, ELE intends to promote a culture of valuing all and any LRs, especially for less-resourced languages, judiciously balancing the importance given to the size and quantity of the LRs and their quality. For the purposes of the DLE metric, we are currently inclined to apply penalties to very small LRs, but this will be eventually confirmed in the final full specification of the DLE concept.

3.2. Year and Contemporaneity of Language Resources

For many applications and NLP tasks, recent and up-to-date LRs are preferable over older ones. While the latter still have value for a range of purposes, all else being equal, the year of collection, production, release or publication of LRs (and, to a certain extent, also LTs) will be considered in the DLE metric; a more recent and more contemporary resource will receive a higher score over a similar older resource, on the assumption that being more current reflects more accurately and comprehensively contemporary language use. The resource is, therefore, more likely appropriate for use by the most recent and advanced methodologies.

[8] http://www.elri-project.eu
[9] https://lr-coordination.eu
[10] https://curlicat-project.eu
[11] http://principleproject.eu

It is important to distinguish between the year of production, completion or publication of the resource on the one hand, and the temporal coverage of the language use it represents overall on the other, as the latter refers to another dimension (e. g., in the case of diachronic corpora, (say) containing newspaper articles from past decades).

3.3. Maintenance and Updates

For the purposes of the ELE Project and the overall concept of DLE, versions and updates are desirable, due to the importance of the criterion of timeliness and contemporaneity of LRTs (Section 3.2). The approach adopted in ELE is to give clear guidelines (Giagkou and Piperidis, 2021) to contributors who are in charge of collecting LRs, also to account for different versions and updates. One way to ensure that this requirement is met is to give full control, sense of ownership and full responsibility to the language experts who will be populating the ELG database as part of the ELE project (almost all of them are ELG National Competence Centre leads, ELRC National Anchor Points, etc., so already intimately familiar with the field and its specifics for the respective language).[12] Anything relevant for a particular language should be run past the team of language experts, who are responsible for collecting the language technologies, language resources, projects and organisations, and curating the resources with respect to what is already registered in the ELG. To ensure that there is one single point of entry into the ELG, for languages spoken in multiple countries, there will be a need to assign one leading expert (e. g., one for German in Germany, Austria and Switzerland; one for French in France, Belgium, Luxembourg; one for Dutch/Flemish, etc).

3.4. Cost, Licensing and Access

The ELE project takes a very inclusive approach to collecting all and any LRTs, irrespective of their cost, licence and access conditions, in the sense that these are documented and recorded, and will be considered when computing the DLE metric. The thinking behind this decision is that a (large) resource that has been tested or evaluated after being built professionally should be recorded in ELG even if it has a cost (for purchase, licensing, use, conditioned upon membership of a particular group or organization, etc.); such licensing restrictions and costs associated with accessing and using a LRT will incur proportional 'penalties' in the DLE metric (Deliverable D1.3).

The aspect of accessibility of LRTs also relates to the broad and varied class of licences often called 'open source'. In reality, especially when it comes to commercial use and exploitation, the different open source licenses differ significantly. For example, a very large data set that is licensed as Creative Commons Zero (CC0) can be used commercially while a different very large data set licensed as Creative Commons Attribution Non-Commercial (CC-BY-NC) cannot. Accordingly, these two data sets would contribute to DLE differently, i. e., the former data set potentially has a bigger impact, even allowing for commercial use and productization, while the existence of the latter is important for research use. One approach to tackle this scenario could be to get the applicable licensing terms (at least the main ones) inspected by legal experts trained in copyright, licensing and intellectual property issues, and placed on a scale ranging from the widest possible licensing terms to the most restrictive: while the most permissive terms will be rewarded for the score of the DLE metric, increasingly restrictive licensing terms will be progressively penalized; this method may also be applied to custom contributor-defined licensing terms, following inspection by a legal expert acting in an advisory capacity. While we recognize that this approach is quite resource-intensive, we intend to keep an open mind about these issues, which will be carefully revisited in preparation

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

for Deliverable D1.3, to investigate the actual feasibility within the project and its eventual follow-up initiatives.

4. Conclusions and Next Steps

This deliverable presents the preliminary definition of DLE, based on a set of modular quantifiers, measures or indicators, along with the DLE metric. The DLE metric can be used to accurately reflect the level of support of language technologies for European languages as an essential requirement of DLE for the present as well as for the future, especially to promote the achievement of full DLE in Europe by 2030 (see Figure 2). This deliverable will be followed up by D1.3 "Digital Language Equality – full specification of the concept" in project month 13 (January 2022), which will present a working, operational full specification of DLE based on well-defined quantifiers, measures, and indicators that should possess descriptive, diagnostic and predictive value to promote DLE for all European languages, by encouraging the levelling up of LT support where this is specifically required. Deliverable D1.3 will include the scoring and weighting mechanisms of the technological and contextual factors introduced here.

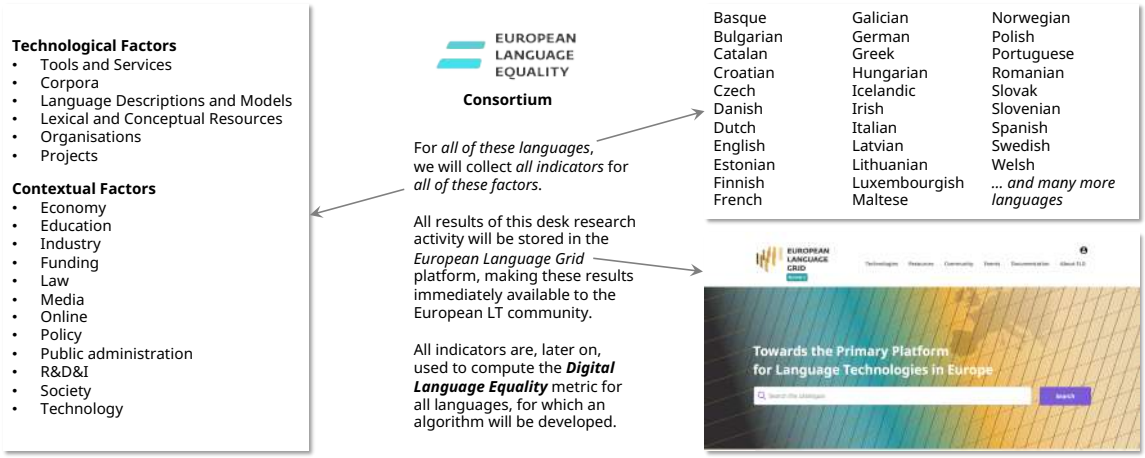

Figure 2: Computing the Digital Language Equality metric

References

Nicoletta Calzolari, Nuria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini Jan Odijk, Stelios Piperidis, Valeria Quochi, and Claudia Soria. *Final FLaReNet Deliverable Language Resources for the Future – The Future of Language Resources The Strategic Language Resource Agenda*. 2011. URL www.flarenet.eu/sites/default/files/FLaReNet_Book.pdf.

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

Maria Giagkou and Stelios Piperidis. *Guidelines for Task 1.3 contributors - Internal working document of the European Language Equality (ELE) project*. 2021.

Steven Krauwer. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, Moscow, Russia, 2003.

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

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 and Hans Uszkoreit, editors. *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer, Heidelberg, New York, Dordrecht, London, 2013. URL http://www.meta-net.eu/sra. More than 200 contributors from research and industry.

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, 5 2020. European Language Resources Association (ELRA).

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

Appendices

A. Technological Factors

Table 1: Digital Language Equality – Technological factors

Category	Factor
Tools and Services	• Language(s) • Domain(s) • Creation/publication date • Licence • Technology Readiness Level • Type of access • Function(s) / Task(s)[13] • Language dependent • Language(s) of output • Media type(s) of input • Media type(s) of output
Corpora	• Language(s) • Domain(s) • Creation/publication date • Licence • Type of access • Annotation type • Corpus subclass • Media type(s) of parts • Multilinguality type • Corpus size, based on corpus size unit

Continued on next page

[13] This factor, and various others, will be aligned with the ontology used in the ELG metadata scheme.

Table 1 – *Continued from previous page*

Category	Factor
Language Descriptions and Models	• Language(s) • Domain(s) • Creation/publication date • Licence • Subclass of grammar/model
Lexical and Conceptual Resources	• Language(s) • Domain(s) • Creation/publication date • Licence • Lexical/conceptual resource subclass • Media type(s) of parts • Encoding level • Number of entries (size)
Projects	• Language(s) of interest • Technology sectors, areas, specialties • Domains (if any) • Duration (based on start and end dates) • Budget • Overall person months
Organizations	• Type: research centre, higher education institution, company, NGO, think tank, public administration • Language(s) of interest • Technology sectors, areas, specialisms • Domains (if any) • Number of people working in the organization • Number of individual members • Number of corporate/institutional members

B. Contextual Factors

Table 2: Digital Language Equality – Contextual factors

Category	Factor
Economy	• Size of the economy of the respective country, countries, region(s) • Size of the LT/NLP market in the respective country, countries, region(s) • Size of the language service and translation or interpreting market in the respective country, countries or region(s) • Percentage of the IT/ICT sector relative to the whole economy of the respective country, countries or region(s) • Investment instruments or accelerator programs targeting AI/LT/NLP start-ups • Regional or national LT/NLP/LSP etc. market (including forecast) • Average socio-economic status of members of the language community
Education	• Number of Higher Education Institutions operating in the language • Percentage of higher education conducted in the language (vs. in English) • Number of academic positions in AI, LT, NLP, computational linguistics, corpus linguistics, language learning/teaching and digital technology, applied linguistics, etc. in the respective country, countries or region(s) • Number of academic programmes of study in AI, LT, NLP, computational linguistics, corpus linguistics, language learning/teaching and digital technology, applied linguistics, etc. in the respective country, countries or region(s) • Literacy level for the language in question • Number of students in language/LT/NLP curricula • Equity in education and educational outcomes • Inclusion in education

Continued on next page

Table 2 – *Continued from previous page*

Category	Factor
Funding	
	• Amount of public funding available for LT/NLP/AI research projects (average or total over a certain number of years) • Venture capital available in the respective country, countries or region(s) • Amount of public funding for interoperable platforms and research infrastructures in the field
Industry	
	• Number of companies developing LTs in or for the respective language • Overall number of start-ups per year (average over a certain number of years) • Specific number of start-ups in the areas of LT/AI/NLP/NLU, etc. (average over a certain number of years)
Law	
	• Copyright legislation and regulations • Legal status and legal protection of the language
Media	
	• Amount of publicly available manually subtitled or dubbed films, tv programmes, online videos, etc. in the language • Amount of publicly available manually transcribed podcasts in the language

Continued on next page

Table 2 – *Continued from previous page*

Category	Factor
Online	
	• Number of digital libraries for the language • Impact of language barriers on e-commerce or other horizontal sectors or domains • Level of digital literacy of members of the language community • Number or size of wikipedia pages for the language (e. g., in comparison to English wikipedia pages) • Number of websites with content available exclusively in the language • Number of websites with content available in the language (but not exclusively) • Number of web pages in the language • Ranking of websites delivering content in the language[14] • Number of labels and lemmas for the language in large public knowledge bases such as Wikidata[15] • Language support gaps according to World Wide Web Consortium (W3C)[16] • Number of ecommerce websites or web shops offering services in the language
Policy	
	• Presence of local, regional or national strategic plans, agendas, committees working on the language, LT, NLP, etc. • Level of recognition and promotion of the LR ecosystem by national or regional authorities • Consideration of regional or national bodies for the citation of LRs in research activities • Promotion of regional, national or international cooperation by the authorities • Level of public and community support for the definition and dissemination of resource production best practices, e. g., enforcing recycling, reusing and repurposing • Existence of policies to provide, maintain and update Basic Language Resources Kits (BLARKs)

Continued on next page

[14] For example, with regard to a well-known ranking such as https://www.alexa.com.
[15] https://multilingual.com/issues/sept-oct-2019/wikidata-gets-wordier/
[16] https://www.w3.org/blog/news/archives/8913

Table 2 – *Continued from previous page*

Category	Factor
Public administration	• Languages of public institutions in the country, countries or region(s) • Number of public services offering services in the language of interest
Research & Development & Innovation	• Innovation capacity (e. g., based on the Innovation Scoreboard position or comparable metric of the respective country, countries or region(s)) • Number of LT, AI, NLP, NLU etc. research groups in total • Number of LT, AI, NLP, NLU etc. research groups or companies predominantly working on the respective language (instead of, say, English) • Overall number of Research & Development staff involved in LT/NLP/NLU(-related), etc. activities • Suitably trained and qualified Research & Development staff (e. g., at doctoral level) in the areas of Number of LT, AI, NLP, NLU etc. in a given time period (e. g., one year) • Capacity for talent retention in the areas of Number of LT, AI, NLP, NLU • State of play of NLP/AI at large when it comes to language understanding • Number of scientists and researchers working on the language (in the different related fields: linguistics, CS, LT, AI, etc.) • Number of researchers and scholars whose work benefits from the availability of or access to language resources, tools and technologies in or for the language • Overall research support staff • Scientific associations or general scientific and technology ecosystem for the language • Number of papers in major conferences and journals reporting studies on language (average over a certain number of years)

Continued on next page

Table 2 – *Continued from previous page*

Category	Factor
Society	
	• Importance, relevance or recognition of the language in the digital age in the respective country, countries, region(s), language community or communities • Number or proportion of fully proficient (literate) speakers of the language • Number or proportion of speaker population with digital skills • Overall number of speakers of the language • Percentage of population that does not speak the official language(s) of the country, region or community, on the basis of socio-demographic factors such as age-group, level of education, income band. • Number of official languages and recognised minority and regional languages in the country, region or community • Number of community languages in the country, countries, region(s) and percentages spoken by the population • Available time resources of the members of the language community • Number of civil society stakeholders working on (preserving) the respective language • Speakers' (positive/negative) attitudes towards the language (e. g., vs. their attitudes towards English) • Involvement of indigenous peoples, particularly women and youth through their own governance structures and representative bodies to support indigenous languages, respecting multiculturalism, ethical standards and integrating the values of indigenous peoples as a form of empowerment. • Sensitivity to barriers that impede the availability of new technology, content and services to indigenous language users • Number or proportion of speaker population who use social media and social networks in the language
Technology	
	• Presence or percentage of open-source language technology • Access to computer, smartphone etc. of members of the language community • Digital connectivity and Internet access in the country, countries, region(s), language community or communities