EUROPEAN LANGUAGE EQUALITY

D3.4

Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap

Authors	ELE Consortium
Dissemination level	Public
Date	18-11-2022

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	D3.4
Deliverable title	Strategic agenda including roadmap
Type	Report
Number of pages	106
Status and version	Version 1.0
Dissemination level	Public
Date	18-11-2022
Work package	WP3: Development of the Strategic Agenda and Roadmap
Task	Task 3.2 Consolidation and aggregation of all input received
Authors	ELE Consortium
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

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
AI4EU	AI4EU (EU project, 2019-2021)
ALPAC	Automatic Language Processing Advisory Committee
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
BLOOM	World's Largest Open Multilingual Language Model
CAGR	Compound annual Growth Rate
CF	Contextual Factor
CEF AT	Connecting Europe Facility, Automated Translation
CELT	Centre of Excellence for Language Technology
CH	Cultural Heritage
CITIA	Conversational Interaction Technology Innovation Alliance
CLAIRE	Confederation of Laboratories for AI Research in Europe
CLARIN	Common Language Resources and Technology Infrastructure
CLTL	Cross-lingual Transfer Learning
CRACKER	Cracking the Language Barrier (EU project, 2015–2017)
CULT	European Parliament's Committee on Culture and Education
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DL	Deep Learning
DLE	Digital Language Equality
DH	Digital Humanities
DNLU	Deep Natural Language Understanding
DSM	Digital Single Market
EBLUL	European Bureau for Lesser Used Languages'
EC	European Commission
ECRML	European Charter for Regional or Minority Languages
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
ELEXIS	European Lexicographic Infrastructure
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
ELRC-SHARE	Repository for Language Resources
EOSC	European Open Science Cloud
EP	European Parliament
ESFRI	European Strategy Forum on Research Infrastructures
EUDAT	European Data (Infrastructure)
EURAMIS	European advanced multilingual information system
euRobotics	euRobotics AISBL
FAIR Principles	Findability, Accessibility, Interoperability, and Reuse
GA	Grant Agreement
GA	General Assembly (ELE project management)
Gaia-X	A Federated Secure Data Infrastructure
GDPR	General Data Protection Regulation
GPT-3	Generative Pre-trained Transformer 3
GPU	Graphics Processing Unit
HCI	Human Computer Interaction (see HMI)
HMI	Human Machine Interaction (see HCI)

HPC	High-Performance Computing
IATE	Interactive Terminology for Europe
ICT	Information and Communications Technology
IoT	Internet of Things
IPR	Intellectual Property Rights
ITRE	European Parliament's Committee on Industry, Research and Energy
LIBER	Europe's principle association of research libraries
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LR	Language Resources/Resources
LRTs	Language Resources and Technology
LT	Language Technology/Technologies
LTC	European Language Technology Council
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
META-SHARE	EU Network of Repositories
MDSM	Multilingual Digital Single Market
ML	Machine Learning
MLLMs	Multilingual Language Models
MT	Machine Translation
NCC	National Competence Centre
NEM	The New European Media18 Initiative
NGO	Non-governmental organization
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PMO	Project Management Office
RDA	Research Data Alliance
RDF	Resource Description Framework
RI	Research Infrastructures
RML	Regional and Minority Languages
R&D&I	Research, Development and Innovation
SC	Steering Committee (ELE project management)
SCIC	Speech Repository
SR	Speech Recognition
SRA	Strategic Research Agenda
SRIA	Strategic Research and Innovation Agenda
SSH	Social Sciences and the Humanities
SSHOC	Social Sciences and the Humanities Open Cloud
STOA	Science and Technology Options Assessment
TF	Technological Factor
TRL	Technology Readiness Level
VR/AR	Virtual Reality/Augmented Reality
WER	Word Error Rate

Executive Summary

The overall vision of the ELE Programme is to achieve complete digital language equality in Europe by 2030. The programme was prepared jointly with many relevant stakeholders from the European Language Technology (LT), Natural Language Processing (NLP), Computational Linguistics and language-centric AI communities, as well as with representatives of relevant initiatives and associations, and language communities. The ELE Programme responds to the call "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", as specified by the European Parliament Resolution *Language equality in the digital age* in 2018. The results of the ELE project show that English is still by far the language with the best and most thorough technological support, followed by a cluster of three languages (German, Spanish, French) that already have only half the technological support of English. After yet another gap, the long tail of languages with fragmentary support starts with Finnish, Italian and Portuguese. More than half of the approx. 90 languages surveyed have either weak or no technological support at all. In comparison to previous results from 2012, the gap between English and the other languages appears to be getting *bigger* instead of smaller. With the exceptions of English, German and French, all languages we investigated exist in socio-political and economical ecosystems that do *not* incentivise, encourage or foster the development of technologies for these languages. While all 30 European countries we surveyed have put in place national AI strategies, almost all of these national strategies seem to have either ignored or left out the topic of languages and language-centric Artificial Intelligence.

The ELE Programme is foreseen to be a shared, long-term, coordinated and collaborative Language Technology funding programme tailored to Europe's needs, demands and values – among others, multilingualism and language equality in general. For the EU we foresee the role of providing resources for coordinating the programme, for providing shared infrastructures, for maintaining the scientific goals and programme principles etc. The participating countries have the role of providing resources for the development of technologies and datasets for their own languages. Key goals are to reduce the technology gap between English and all other European languages and to address the lack of available language data – this is true for all European languages except English. The ELE Programme focuses upon *openness*: open source, open access and open standards as well as interoperability and standardisation. It makes use of and strengthens existing as well as emerging infrastructures and data spaces. With regard to the scientific dimension, the ELE Programme attempts to achieve the goal of *Deep Natural Language Understanding by 2030*. A key emphasis is on the creation of large open access language models for all European languages including the creation of datasets, multilingual models, models that include symbolic knowledge, models that include discourse capabilities as well as grounding and other sophisticated features currently out of reach for existing state of the art technologies. The ELE Programme is foreseen to have a runtime of nine years, divided into three phases of three years each. In addition to the overall coordination, the ELE Programme tackles the following overarching themes: *Language Modelling*, *Data and Knowledge*, *Machine Translation*, *Text Understanding* and *Speech*. All of these interconnected themes focus upon the socio-political goal of establishing digital language equality in Europe and on the scientific goal of Deep Natural Language Understanding, both by 2030. The ELE Programme is designed in such a way that it makes optimal use of infrastructures and services developed in relevant other European initiatives.

The global NLP market is estimated to reach 341.7B$ by 2030. In contrast, the modest investment needed to implement the ELE Programme will not only bring about digital language equality in Europe, it will also move European research and industry in this field into a dominating position for years to come.

1 Multilingual Europe and Digital Technologies

In varietate concordia (in English: *united in diversity*[1]) is the official Latin motto of the EU, adopted in 2000. According to the European Commission,

> The motto means that, via the EU, Europeans are united in working together for peace and prosperity, and that the many different cultures, traditions and **languages in Europe** are a positive asset for the continent.[2] [*emphasis added*]

In Europe's multilingual setup, all 24 official EU languages are granted equal status by the EU Charter and the Treaty on the EU; moreover, the EU is home to over 60 regional and minority languages which are protected and promoted under the European Charter for Regional or Minority Languages (ECRML) treaty since 1992,[3] in addition to migrant languages and various sign languages, spoken by some 50 million people. Furthermore, the Charter of Fundamental Rights of the EU under Article 21[4] states that,

> Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, **language**, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited. [*emphasis added*]

However, language barriers still hamper cross-lingual communication and the free flow of knowledge and thought across languages. Multilingualism is one of the key cultural cornerstones of Europe and signifies part of what it means to be and to feel European. However, no common EU policy has been proposed to address the problem of language barriers.

To help repair the economic and social damage caused by the pandemic, the EU has agreed on a recovery plan to lead the way out of the crisis towards a modern and more sustainable Europe. The EU's long-term budget for 2021-2027, coupled with NextGenerationEU, the temporary instrument designed to boost the recovery, will be the largest stimulus package ever financed through the EU budget. A total of €1.8 trillion will help rebuild a post-COVID-19 Europe.[5] NextGenerationEU is a €750 billion temporary recovery instrument to help repair the immediate economic and social damage brought about by the coronavirus pandemic. More than 50% of the amount will support modernisation, for example through research and innovation, via Horizon Europe and the digital transition, via the Digital Europe Programme.[6]

The European Language Technology (LT) community is committed to do the research, development and deployment of ground-breaking and novel technologies in order to successfully turn a linguistically fragmented Europe into a truly unified and inclusive one. By fully supporting the rich and diverse linguistic cultural heritage from broadly spoken languages to minority and regional languages as well as the languages of immigrants and important trade partners, it will benefit the European citizen, European industry and European society.

Europe is in need of powerful LT made in Europe for Europe and all European citizens, tailored to its specific cultures and societies as well as economic demands. While language diversity is at the core of Europe's identity and multilingual society, many languages are in danger of digital extinction because they are not sufficiently supported through LT. The META-NET White Paper Series (Rehm and Uszkoreit, 2012) and its follow-up language reports have revealed that there is a steadily increasing and rather severe threat of digital extinction for most European languages. The study "Language Equality in the Digital Age – Towards a

1 https://europa.eu/european-union/about-eu/symbols/motto_en
2 http://europa.eu/abc/symbols/motto/index_en.htm
3 https://en.m.wikipedia.org/wiki/European_Charter_for_Regional_or_Minority_Languages
4 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT
5 https://ec.europa.eu/info/strategy/recovery-plan-europe_en
6 https://digital-strategy.ec.europa.eu/en/activities/digital-programme

Human Language Project", commissioned by the European Parliament's Science and Technology Options Assessment Committee (STOA), recommended to initiate a new, large-scale European LT research, development and innovation programme in response to this threat. Shortly after the publication of this study in 2018, the European Parliament adopted with a landslide majority of 592 votes in favour a resolution on "language equality in the digital age" that also included the suggestion to intensify research and funding to achieve deep natural language understanding. Multilingualism is under the scope of a series of EU policy areas, including culture, education, the economy, the Digital Single Market (DSM), lifelong learning, employment, social inclusion, competitiveness, youth, civil society, mobility, research and media. More attention needs to be paid to removing barriers to intercultural and interlinguistic dialogue, and to stimulate mutual understanding (Rehm and Uszkoreit, 2012; STOA, 2017; European Parliament, 2018). Sophisticated multilingual, crosslingual and monolingual technologies for all European languages would future-proof our languages as cornerstones of our cultural heritage and richness

In recent years, European research in LT has been facing increased competition from other continents, especially with regard to breakthroughs in Artificial Intelligence (AI). These scientific breakthroughs have not only led to global commercial successes, but also encouraged European scientists, including young high achievers, to leave Europe and continue their research abroad.

The European LT community is committed to providing robust and novel technologies in order to successfully turn a fragmented environment into a truly unified and inclusive Europe, supporting our rich and diverse linguistic heritage. Thereby, European LT should foster and support multilingualism while strictly adhering to European values such as privacy by design, transferability, fairness, diversity and openness, transparency and accountability, public wealth, individual rights and collective purposes.

Recognizing that European languages are currently not equally served and supported in terms of digital services and opportunities will encourage the development of technologies, tools and resources that at present are available only for a small number of thriving languages.

1.1 Europe's Languages in the Digital Sphere

Natural language is at the heart of human intelligence.

Languages are the most common and versatile way for humans to convey and access information. We use language, our natural means of communication, to encode, store, transmit, share and manipulate information. We use language in everyday life, to interact with others and our environment and as social glue, to express and to explain ourselves, to convince, agree with, rebut others etc. Our laws and constitutions are written in language. We use it in science, commerce, in teaching and passing on knowledge to the next generations, for pleasure, creativity and aesthetic enjoyment in puns, jokes and art. History and culture are recorded, interpreted and enjoyed through language. Our languages are a core part of our identities.

Human languages are incredibly complex: a single word (phrase, sentence, text) can have many meanings, a single meaning can be expressed by many different words (but meaning depends on linguistic and situational context), we can use language literally and metaphorically, language and knowledge are highly intertwined, we do not articulate important parts of a message if these parts are presumed shared knowledge by the community (this includes situational knowledge), important parts of meaning reside in what can be inferred from what has been said etc. At the same time, language changes: new words are invented, some old ones are dropped, even the structure (syntax and morphology) of languages and the meaning of words change over time. These aspects make human languages fundamentally

different from the formal languages of mathematics, logic and computer science. This is also what makes human languages so efficient, elegant, flexible and enjoyable. Finally, there are many human languages (6000+), not even counting regional and dialectical variants. All these aspects are at the core of human languages and they make it hard for computers to "fully understand" human language and to "properly" process human language in the context of "full and deep understanding".

Languages are at the heart of every aspect of life and their role is crucial to the future of European countries, citizens, businesses, and of the Union as a whole. Language equality (Gaspari et al., 2022b) can deliver impact in the following four high-priority areas.

- **Language equality** will have a positive and unprecedented impact on under resourced European languages. It needs to be ensured that no European languages remain under-resourced, but that they can be equipped with the same very high level of technological support already enjoyed by very few of them. This, in itself, will deliver major impact on all European citizens and businesses: supporting all languages in the interest of equality and fairness empowers and brings advantages to their speakers, while reflecting the democratic and inclusive spirit of the EU.
- **Language equality** will make a contribution to establishing a fair, inclusive and sustainable multilingual DSM: this will be achieved by helping to make all European languages future-proof through digital technologies, and especially preventing the threat of digital extinction for the ones that suffer from chronic weak support. By fostering a more inclusive and cooperative business and social environment, companies and citizens will benefit from sharing knowledge, digital services and products on an equal footing, overcoming the fragmentation that is caused by several European languages lagging behind, which severely penalizes their speakers as well as regional and local communities. Action in this vital area is particularly urgent due to the increasing range of economic, educational and social opportunities that are afforded online and delivered remotely, from e-commerce to online shopping, to web-based recruitment services, online teaching programmes and professional training courses, etc.
- **Language equality** will help science and research in Europe, mobilising and leveraging their full potential to start reclaiming scientific and industrial leadership from US-based and Asian competitors, particularly tech giants as well as academic institutions and research centres, that pose fierce competition in several research-led fields. It will instigate regional, national and EU-wide collaboration among scientists from academia and industry covering a broad range of disciplines, ensuring the mix of competencies that is required to deliver substantial impact at the forefront of scientific and technological advancement.
- **Language equality** will act as a multiplier of opportunities. It will help to aggregate the players that are required to unlock the full potential of an EU-wide effort to exchange and share widely-agreed methodologies, resources and technologies with a focus on promoting the digital equality of European languages: this will benefit the use and promotion of all European languages, encouraging in particular those that have traditionally lagged behind.

1.2 What is Language Technology and how can it help?

Language Technology (LT) is concerned with studying and developing systems capable of processing human language. Over the years, the field has developed different methods to make the information contained in written and spoken language – and increasingly for other modalities such as sign language, for example – explicit or to generate or synthesise written

or spoken language (see Section 3 for more detail). Despite the inherent difficulty of many of the tasks performed, current LT support allows many advanced applications which have been unthinkable only a few years ago. Language Technology is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, text editors, text predictors, automatic translation systems, automatic subtitling, automatic summaries, inclusive technology, etc. Its rapid development in recent years predicts even more encouraging and also exciting results in the near future.

Language Technology is providing solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e. the generation of speech given a piece of text, Automatic Speech Recognition, i. e. the conversion of speech signal into text, and Speaker Recognition (SR).
- **Machine Translation** i. e. the automatic and assistive technologies to help translating from one natural language (including sign languages) into another.
- **Information Extraction** and **Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.
- **Natural Language Generation** (NLG) is the task of automatically generating texts. Summarisation, i. e. the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction** aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). A very popular application within this area are conversational agents (better known as chatbots).

1.3 Language Technology and Artificial Intelligence

LT is at the heart of the software that processes unstructured information and exploits the vast amount of data contained in text, audio and video files including those from the web, social media, etc. Only the proper application of LT will allow processing and understanding, i. e., making sense of these enormous volumes of multilingual written and spoken data in sectors as diverse as health, justice, education, or finance. LT applications such as speech recognition, speech synthesis, textual analysis and machine translation are actually used by hundreds of millions of users on a daily basis. As reflected in the European, national and regional AI and LT strategies both inside and outside Europe (Aldabe et al., 2021a), LT is outlined as one of the most relevant technologies for society, as seen by its inclusion in all the prioritized strategic areas for developing research, development and innovation (R&D&I) activities. LT is multidisciplinary in nature since it combines knowledge in computer science (and specifically in AI), mathematics, linguistics and psychology among others. This uniqueness must be considered in any public or private initiative in AI that includes LT.

In recent years, the LT community is contributing to the emergence of powerful new deep learning techniques and tools that are revolutionizing the approach to LT tasks. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The current success in several areas of AI is possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches (Goodfellow et al., 2016; Devlin et al., 2019; Liu et al., 2020; Torfi et al., 2020; Wolf et al., 2020).

These international reports, in addition to several others, prepared between 2018 and 2022, place LT as one of the three most important functional application areas within AI together with Vision and Robotics. Automatic language understanding is perceived as one of the fundamental goals of AI, and, in turn, it is also considered one of its main challenges.

1.4 The European Language Technology Community

The success of the European LT community critically depends on the close collaboration and cooperation with many different stakeholders, most importantly in industry and research, but also in administration, politics, civil society and standardisation as well as on powerful instruments for informing and mobilising stakeholders on the regional, national and international level.

1.4.1 The European Language Technology Community: Research

Europe has a long-standing research, development and innovation tradition in LT with over 800 centres performing excellent, highly visible and internationally recognised research on all European and many non-European languages.

Research centres, universities and other academic institutions that do research in Language Technology, Computational Linguistics, Language-centric AI, Knowledge Technologies, Cognitive Science, Linguistics etc. form one important branch of the Language Technology Community.

Founded in 2010, META-NET[7] is a European Network of Excellence dedicated to the technological foundations of a multilingual and inclusive European society, bringing together 60 research centres in 34 European countries. One of its main goals is technology support for all European languages as well as fostering innovative research by providing strategic recommendations with regard to key research topics (Rehm and Uszkoreit, 2013). META-NET inspired META-SHARE,[8] an infrastructure bringing together providers and consumers of language data, tools and services and providing a network of repositories that store resources, documented with high-quality metadata aggregated in central inventories (Gavrilidou et al., 2012; Piperidis et al., 2014).

CLARIN (Common LAnguage Resources and technology INfrastructure) is one of the pan-European research infrastructures (RIs) that form the RI landscape that is supported and monitored by ESFRI.[9] It is strongly rooted in the humanities and the field of NLP and has the mission to create and maintain an infrastructure to support the sharing, use and sustainable availability of language data and tools for research in the Social Sciences and Humanities (SSH) and beyond.[10] Since its early days, the CLARIN consortium has aimed at building both

[7] http://www.meta-net.eu
[8] http://www.meta-share.org
[9] https://www.esfri.eu/esfri-roadmap-2021
[10] See https://www.clarin.eu/content/vision-and-strategy

a technical infrastructure and a sustainable organisation for collaboration and coordination across the participating national consortia, as well as the exchange of knowledge and best practices, (see e.g. Broeder et al. (2008); Hinrichs and Krauwer (2014)). The CLARIN infrastructure adheres to the interoperability paradigm on several levels, including metadata harmonisation and standardisation (de Jong et al., 2020). The CLARIN consortium was established as a legal entity in 2012. It is a so-called ERIC (European Research Infrastructure Consortium), which is based on a model for funding and governance by the participating parties, with room for in-kind contributions from national consortia and independent third parties, both from Europe and beyond.

The Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE)[11] is an organisation created by the European AI community that seeks to strengthen European excellence in AI research and innovation, with a strong focus on human-centred AI. CLAIRE aims to ensure that societies and citizens across all of Europe, and beyond, benefit from AI as a major driver of innovation, future growth and competitiveness, and to achieve world-wide brand recognition for "AI made in Europe".

Founded in 2018, CLAIRE has garnered the support of more than 3,700 AI experts and stakeholders, who jointly represent the vast majority of Europe's AI community, spanning academia and industry, research and innovation. Among them are more than 140 fellows from various key scientific associations. CLAIRE's membership network consists of over 430 research groups and institutions, covering jointly more than 24,000 employees in 37 countries. Furthermore, CLAIRE has recently set up an Innovation Network that, together with the established Research Network, will foster a strong link between research and industry.

CLAIRE strongly believes that LT and NLP play a key role not only in Europe but also around the world. CLAIRE supports the mission to leverage the capabilities and potential in that area for the benefit of everyone. According to a survey recently conducted among the CLAIRE member groups, 44.9% indicated to have (some) expertise in NLP. CLAIRE has an Advisory Group on NLP[12] consisting of high-caliber European NLP and LT scientists for advice on the needs of the community and on how to connect to it.

1.4.2 The European Language Technology Community: Industry

The European LT industry has been estimated to comprise 435 companies, according to LT-Innovate (2016) or 473 LT vendors in EU26 plus Iceland and Norway in 2017 (Vasiljevs et al., 2019).

The European Language Grid (ELG)[13] is a direct follow-up project to META-NET. The data sets, resources, models, tools from META-SHARE and other initiatives such as ELRC-SHARE and ELRA have been added to the ELG catalogue.

The ELG cloud platform is targeted to evolve into the primary platform for Language Technology in Europe. Its aim is to provide one umbrella platform for the Language Technology developed by the European LT community, including research and industry, addressing a gap that has been repeatedly raised by the European Parliament (STOA, 2017; European Parliament, 2018) and by the European LT community in a number of strategy papers throughout the years (Rehm and Uszkoreit, 2013; Rehm et al., 2016b; Rehm, 2017; Rehm and Hegele, 2018; Rehm et al., 2020b,a).

The ELG is meant to be a virtual home and marketplace for all products, services and organisations active in the LT space in Europe (Rehm et al., 2020a). It enables the European LT community to deposit and upload their technologies and data sets and to deploy them through the grid. The platform can be used by all stakeholders to showcase, share and dis-

[11] https://claire-ai.org
[12] https://claire-ai.org/iags
[13] https://www.european-language-grid.eu

tribute their products, services, tools and resources. At the point of writing, the ELG is still funded by the EU (2019-2022); it will establish a legal entity in the first half of 2022, so that the platform can continue to provide access to the commercial and non commercial tools and services as well as language resources it hosts.

In a wider context, the ELG is also meant to support digital language equality in Europe (STOA, 2017; European Parliament, 2018), i. e., to create a situation in which all languages are supported through technologies equally well. The current imbalance is characterised by a stark predominance of language resources for English, while almost all other languages are only marginally supported and, thus, in danger of digital language extinction (Rehm and Uszkoreit, 2012; Kornai, 2013; Rehm et al., 2014, 2016a; Berzins et al., 2019a).

In October 2022, the ELG catalogue comprises more than 880 commercial entities, also including integrators and a certain number of user companies (Rehm et al., 2020a, 2021).

1.4.3 Users of European Language Technology

Users of European LT form a diverse target group that can include virtually everyone in Europe. Language Technology is nowadays used by very large segments of the European population, often even unconsciously. Despite the fact that this group is heterogeneous, sectors and industries like media and broadcasting networks, healthcare, banking and insurance, e-commerce, mobility, telecommunications or public administrations have been identified to profit immensely.

Many of these industry and public stakeholder groups would highly benefit from LT systems but do not have access to it. Relevant technologies to be explored include LT applications that are specific to a work environment: customer interaction technologies in business and trade, educational applications, e. g. for language training, documentation and support systems in hospitals and care facilities or chatbots for queries in administrations on local, regional and national levels, to name a few.

Language communities include all speakers of Europe's languages, essentially all European citizens. Different language communities have different needs, but especially communities with smaller numbers of speakers rely on the support of these federations. Here, the notion of trust again plays a crucial role. These representative bodies give a voice to communities that would otherwise hardly be heard. That having been said, especially these language communities can benefit the most from the mutual goal of establishing digital language equality in Europe. There are numerous related umbrella networks and initiatives that are of importance.

The European Federation of National Institutions for Language[14] (EFNIL) provides a forum for member institutions to exchange information about their work and to gather and publish information about language use and language policy within the European Union. In addition, the Federation encourages the study of the official European languages and a coordinated approach towards mother-tongue and foreign-language learning, as a means of promoting linguistic and cultural diversity within the European Union.

The European Language Equality Network[15] (ELEN) has as its goal the promotion and protection of European lesser-used (i.e. regional, minority, endangered, indigenous, co-official and smaller national) languages, to work towards linguistic equality for these languages, and multilingualism, under the broader framework of human rights, and to be a voice for the speakers of these languages at the local, regional, national, European and international level. ELEN is a non-governmental organisation. ELEN was established in 2011 based on the former European Bureau for Lesser Used Languages' (EBLUL) member-state committees, Eurolang, plus many umbrella and individual language NGOs from most EU member states.

[14] http://www.efnil.org
[15] https://elen.ngo

ELEN's purpose is to represent the 50 million people, 10% of the EU's population, who speak a regional, minority, or endangered language. ELEN represents 44 regional, minority and endangered languages in 18 States, so far.

The European Civil Society Platform for Multilingualism[16] (ECSPM) is an alliance for languages and multilingualism in Europe, making possible the cooperation between European, national, and international networks, organisations, federations and associations that view multilingualism as an asset for European economic, social, cultural development, and as a facilitator for intellectual growth, social, and personal development. It aspires to be a strong voice of Europe's civil society, promoting language policies for multilingualism in all aspects of social life by way of focusing on people, and on their ability to use a variety of semiotic resources to access education, social affairs and culture, to participate as active citizens in the EU, shaping its making, benefiting from better communication, wider employment and business opportunities.

The Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries[17] (LIBER) is the voice of Europe's research library community. Approximately 420 national, university and other libraries are part of LIBER and our wider network includes goal-oriented partnerships with other organisations in Europe and beyond.

The New European Media[18] (NEM) Initiative was established as one of the European Technology Platforms under the Seventh Framework Programme, aiming at fostering the convergence between consumer electronics, broadcasting and telecoms in order to develop the emerging business sector of networked and electronic media. In order to respond to new needs and requirements of the Horizon 2020 programme, the NEM initiative enlarged its focus towards creative industries and changed its name from Networked and Electronic Media Initiative to New European Media, dealing with Connected, Converging and Interactive Media & Industries, driving the future of digital experience.

Wikipedia[19] is a free content, multilingual online encyclopedia written and maintained by a community of volunteers through a model of open collaboration, using a wiki-based editing system. Wikipedia is a project of the Wikimedia Foundation, a non-profit organisation whose aim is to bring knowledge to everyone on the planet. Wikipedia is the best-known project, but Wikimedia offers many other services such as Wiktionary,[20] a free dictionary or Wikidata,[21] which is a collaborative, free and open knowledge base that stores structured information. Its main advantage is that it offers linked data, described using RDF, which allows data to be linked to other datasets in other digital repositories.

1.4.4 Relevant Initiatives

Over the coming years, AI is expected to transform not only every industry, but also society as a whole. While other tasks such as image recognition and robotics have provided testbeds for massive new scientific breakthroughs, LT and NLP are, by now, considered important driving forces. There are several European initiatives that dominate current research and development, many of which have close ties to the ELE project, such as ELG and CLARIN.

The European Commission recently announced that it will set up common European Data Spaces as an integral part of the Digital Europe Programme. The aim of these data spaces is to connect data from various ecosystems and sectors that is currently fragmented and dispersed, while enabling an interoperable and trusted environment for data processing. To extract information from multimodal language data, services trained on large data sets

[16] https://ecspm.org
[17] https://libereurope.eu
[18] https://nem-initiative.org
[19] https://www.wikipedia.org/
[20] https://www.wiktionary.org/
[21] https://www.wikidata.org/wiki/Wikidata:Main_Page

are necessary. The space will be deployed in two work strands. The first will establish an institutional Centre of Excellence for Language Technology (CELT) making use of existing EU initiatives of language data collections such as ELRC, EURAMIS, SCIC repositories, IATE, CLARIN, and META-SHARE. The second will support the deployment of the Language Data Space[22] on the basis of existing EU initiatives such as the European Language Grid and CEF Automated Translation (eTranslation and other LTs) (Support Centre for Data Sharing, 2022).

Another initiative to build a high-performance data infrastructure for Europe which is competitive, secure and trustworthy is Gaia-X.[23] Representatives from business, politics, and science from Europe and around the globe are working together, hand in hand. Companies and citizens will collate and share data – in such a way that they keep control over them. They should decide what happens to their data, where it is stored, and always retain data sovereignty. The architecture of Gaia-X is based on the principle of decentralisation. Gaia-X is the result of a multitude of individual platforms that all follow a common standard – the Gaia-X standard. The result will be a networked system that links many cloud services providers together (Bundesministerium für Wirtschaft und Energie, 2020).

1.5 Market Opportunities

LT is one of the most important AI application areas with a fast growing economic impact. Funding for LT start-ups is booming.[24] Early-stage funding in 2021 amounts to just over USD 1 billion for companies that offer solutions that are based on or make significant use of NLP, providing a picture of what funders think is innovative.[25] Reports from various consulting firms forecast enormous growth in the global LT market based on the explosion of applications observed in recent years and the expected exponential growth in unstructured digital data. For instance, according to an industry report from 2019,[26] the global NLP market size is set to grow from USD 10.2 billion in 2019 to USD 26.4 billion by 2024, at a CAGR of 21.0 percent, during the forecast period 2019-2024.[27] According to another report from the end of 2019,[28] the global NLP market was valued at USD 8.5 billion in 2018, which is expected to reach USD 23.0 billion by 2024, registering a CAGR of 20.0% during the forecast period. A report from 2020 highlights that the global NLP market size stood at USD 8.61 billion in 2018 and is projected to reach USD 80.68 billion at 2026, exhibiting a CAGR of 32.4% during the forecast period.[29] Another report from 2020 estimates the global LT market to reach USD 41 billion by 2025.[30] In a report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at a CAGR of 18.4% from 2020 to 2028.[31] Due to the COVID-19 crisis, the global market for NLP hit already USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at a CAGR of 10.3% over the analysis period 2020-2027 according to a report from 2021.[32] The rated NLP market size for the year

[22] https://digital-strategy.ec.europa.eu/en/funding/language-data-space-call-tenders
[23] https://www.data-infrastructure.eu/
[24] https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/?sh=429aff902b14
[25] https://towardsdatascience.com/nlp-how-to-spend-a-billion-dollars-e0dcdf82ea9f
[26] https://www.businesswire.com/news/home/20191230005197/en/Global-Natural-Language-Processing-NLP-Market-Size
[27] https://www.analyticsinsight.net/potentials-of-nlp-techniques-industry-implementation-and-global-market-outline/
[28] https://www.vynzresearch.com/ict-media/natural-language-processing-nlp-market
[29] https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933
[30] https://www.globenewswire.com/news-release/2020/07/10/2060472/0/en/Natural-Language-Processing-NLP-Market-to-reach-US-41-billion-by-2025-Global-Insights-on-Trends-Leading-Players-Value-Chain-Analysis-Strategic-Initiatives-and-Key-Growth-Opportunit.html
[31] https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html
[32] https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market

2028 suggested by another report exceeds the previous estimated figures by far naming an amount of USD 127.26 billion at a CAGR of 29.4% in the forecasted period.[33] A recent report estimates that NLP specifically in Europe will witness market growth of 19.7% CAGR and is expected to reach USD 35.1 billion by 2026.[34] Finally, another current report rates the size of the NLP market by 2027 as approximately USD 48.46 billion and market growth of 21.3% CAGR.[35]

2 Trends and Mega-Trends in Digital Technologies

There are diverse trends and megatrends that bear closely on digital technologies. Among others, these include accelerating hyperconnectivity, shifts in the nature of work, increasing digitalization, new modes of learning, expanding consumerism, novel approaches to politics and governance, changes in healthcare, and the rapidly evolving field of AI. While the future course of these trends cannot be known with certainty, their current trajectories suggest that LT will play a deciding role in how they unfold. It is therefore beneficial to briefly outline in broad strokes some of the ways digital technologies fit within and shape these trends.

The digital world, still relatively young, is rapidly consolidating into a space that exists alongside our physical reality. As connections deepen and multiply between these two spheres, our digital lives require tools and resources that permit and facilitate interaction with and within the virtual realm. Among these are means to transcend the linguistic barriers that will become evermore problematic as interconnectedness condenses, without forgetting that languages which do not possess sufficient technological support may very well be left behind. LT is uniquely positioned to solve many of the obstacles associated with cross-language digital communication, including those that inhibit the flow and accessibility of information and knowledge across Europe (Gomez-Perez et al., 2022). It is also worth remembering that as individual digital technologies begin to work in unison with greater ease, ambient intelligence will become more pervasive, responsive and natural-feeling. As LT advances, it may soon be difficult to distinguish human-computer from human-human communication, a phenomenon that will help make collaboration with AI commonplace and propel the use of augmented intelligence, such as intelligent personal assistants (Kaltenboeck et al., 2022). By way of example, it is anticipated that half of all knowledge workers will utilize an AI-based virtual assistant on a daily basis by 2025 (up from only 2% in 2019). This development could have economic impacts as well. Gartner predicts AI augmentation will surpass all other types of AI initiatives in terms of business value by 2030.[36]

2.1 Digital Twins and Personal Virtual Worlds

The growing hyperconnectivity, the interaction between data, computers and devices, is frequently catalyzed by the use of AI and sophisticated LT (STOA, 2017; Davis and Philbeck, 2017).[37] Both aid, for instance, in the construction of digital twins, a trend that relies on machine learning to lend decision-making capabilities to virtual replicas of actual objects or

[33] https://www.analyticsinsight.net/the-global-nlp-market-is-predicted-to-reach-us127-26-billion-by-2028/
[34] https://www.analyticsinsight.net/nlp-in-europe-is-expected-to-reach-us35-1-billion-by-2026/
[35] https://www.globenewswire.com/en/news-release/2022/05/25/2450815/0/en/Global-Natural-Language-Processing-Market-is-Expected-to-Represent-a-Value-of-USD-48-46-billion-by-2027-Fior-Markets.html
[36] https://blogs.gartner.com/anthony_bradley/2020/08/10/brace-yourself-for-an-explosion-of-virtual-assistants/ and https://www.gartner.com/en/newsroom/press-releases/2019-08-05-gartner-says-ai-augmentation-will-create-2point9-trillion-of-business-value-in-2021
[37] https://knowledge4policy.ec.europa.eu/accelerating-technological-change-hyperconnectivity_en; https://knowledge4policy.ec.europa.eu/foresight/topic/accelerating-technological-change-hyperconnectivity/developments-forecasts-accelerating-technological-change-hyperconnectivity_en

systems. Utilizing data-driven AI models, digital twins are increasingly prevalent in a variety of social sectors, such as urban planning and domotics, where they can help test city development strategies or manage a household's technical facilities. In a related trend that portends to accelerate, consumers are more willing to spend in order to customize lateral personal virtual worlds. The importance of digital technologies in this area, along with the spread of online personalized shopping experiences, should not be underestimated given that the global middle class may number close to five billion people by 2030, an increase in purchasing power that will influence consumption patterns significantly. A corollary to this development is the belief that AI will add more than €15 trillion to the global economy during this time and that its market valuation will reach €1.4 trillion by 2029, up from €387 billion in 2021 (Backfried et al., 2022).[38] Part of this will come from AI software revenue, which is forecasted to grow significantly.

2.2 LT and the Workplace

Growing digitalization and hyperconnectivity are changing the nature of work, giving rise to new types of employment and a reorganization of business models. The use of professional networks and job databases in employment, for instance, is a trend that is accelerating. These AI-powered services enable employers and job-seekers alike to match CVs with employment opportunities (Gomez-Perez et al., 2022). Presently, work in the digital age is increasingly decentralized and characterized by greater flexibility, while automation and AI are creating demand for digital and knowledge-based skills. It is believed that AI will continue to shape the workplace by increasing productivity and easing work-flow management.[39] In fact, AI is already being utilized by employers as a means to sync projects with employees according to particular skill sets and to assess merit-based promotion. And graph technology for analytics is on the rise in the business world as well. By 2023, graph technologies will facilitate rapid contextualization for decision making in 30% of organizations worldwide. The labor market has yet to feel the full force of these transformations in terms of job loss, but it is possible that automation and AI-enabled machines may eventually lead to disruption in employment. At the same time, a report issued by the IT consulting firm Accenture estimates that AI may double the annual economic growth rates in several developed countries by 2035.[40] Indeed, digital technology has the potential to generate employment and 1.75 million new jobs in the area of ICT are expected to be created by the end of this decade.[41]

2.3 Education and Training

New employment opportunities will require a workforce that is not only skilled in the use and maintenance of digital and AI technologies, but also well-suited for knowledge creation and dissemination in the information economy. If this trend continues to develop, workers may be asked to become proficient at deploying a set of digital skills and AI tools that are common across several professions. Movement in this direction is already underway, as attested to by personalised training programs and platforms, many based on LT, that are in place to create a digitally-skilled workforce. Moreover, increased demand for digital skills and familiarity with AI will put more focus on ensuring they are taught to children, teenagers and young adults as part of school and university curricula. It is likely that digital competence,

[38] https://www.fortunebusinessinsights.com/press-release/artificial-intelligence-market-9227
[39] https://knowledge4policy.ec.europa.eu/foresight/changing-nature-work_en; https://www.wired.com/insights/2013/08/the-rise-of-the-millennial-workforce/; https://knowledge4policy.ec.europa.eu/foresight/topic/changing-nature-work/demographic-trends-of-workforce_en
[40] https://www.accenture.com/_acnmedia/pdf-57/accenture-ai-economic-growth-infographic.pdf
[41] https://knowledge4policy.ec.europa.eu/foresight/topic/changing-nature-work/technological-progress_en

now often garnered through available access to digital technology at home, will become progressively instilled through established digital-literacy educational programs and training. A trend worth mentioning in this regard is the strengthening tie between learning and high tech, a shift that is propelling education towards technologically supported learning (Leahy et al., 2019). AI-powered tutoring systems may become more common as a way to deliver learning material both within and without the classroom. Rather than supplanting teachers, these systems will increasingly function as tools that accompany instructors in an effort to provide more interactive learning and foster critical thinking. One illustration of this are Computer Assisted Language Learning tools that rely on text analysis and natural language understanding. Another is adaptive learning, including virtual assistants or augmented virtual reality, which can personalise instruction by identifying a student's progress in order to tailor a specific curriculum (Gomez-Perez et al., 2022). These strategies are not only capable of gathering learning progress analytics, but can also assist when student-teacher ratios are high or students possess learning difficulties.

2.4 LT and Commerce

The continuing growth in eCommerce within European economies is a trend that is both significant and ripe for LT applications. This includes helping communication across Europe's Digital Single Market, where access to and knowledge about products, national and local policies, trade and finance can be facilitated through LT. Varying content that rapidly changes must be continuously translated across several platforms, including product labeling, social media posts, marketing, and customer opinions. Some of this content may be more technical and some might require cultural context. As interconnectedness strengthens, businesses need to ensure their content is multilingual in order to reach diverse markets effectively. Communicating with customers in their native languages not only helps provide clarity vis-a-vis products and services, but also builds trust between businesses and clients that reside in different European regions (Bērziņš et al., 2022). In a similar fashion, businesses can employ sentiment analysis of social media, reviews and feedback to gauge customer satisfaction and provide better customer service. Along these lines, commercial enterprises are beginning to experiment with the use of data fabric to manage and integrate big data, a trend that is expected to accelerate. This can give businesses a perspective on data that allows them to better understand the interaction between customers and products (Kaltenboeck et al., 2022).

2.5 The Data Marketplace

The increased attention being paid to data fabric is a reminder that the European Commission hopes to utilize the Digital Single Market in conjunction with its European Data Strategy to erect thematic data spaces that will ideally bring the European Union's data economy in line with its economic size. These data marketplaces may also alter how data is shared if they become trusted as spaces that contain transparent data ecosystems bolstered by the guarantee that data is private and anonymous. Such trust is essential given that businesses based on data marketplaces will depend on the willingness of stakeholders to share data. Additionally, the EU data economy will continue to expand over the remainder of the decade and the increasing availability of data may change the nature of data science jobs as demand for data-related positions and knowledge-based skills grows (Simon et al., 2021). The establishment of data marketplaces will give these data professionals the opportunity to become self-employed, offering applications and services directly to buyers.

2.6 Digital Technologies, Government, and Democracy

More time living online has engendered novel perspectives about how government and society might be restructured in the future, including theories on *Do it yourself* democracy, super-collaborative government, and private algocracy (European Commission et al., 2019). It is unlikely any of these scenarios will come to exist as imagined today, but they are worth taking into account when considering how digital technologies could be harnessed. The idea behind private algocracy, for example, envisions a world in which data, data analytics and decision making are in the hands of multinational corporations. In this future, personal data is fully monetized and sole access to Big Data and analytic tools allows these companies to provide private and public services to citizens, including news and other media. As noted, this cautionary sketch reflects one possible trajectory of a rising trend towards new governing systems that is taken to an extreme. However, it is true that digital technologies are altering how government and democracy operate in various ways.[42] Digital technology and AI are increasingly utilized to improve interactions between public administrations and citizens, leading governments to place greater reliance on them. In some cases, this has meant adopting automated decision-making or employing AI to personalize public services and anticipate roadblocks that might arise when crafting public programs. Furthermore, government use of both chatbots and data mining will likely rise significantly over the next few years, including virtual assistants to help citizens locate needed information and programs that facilitate policy analysis. Text analysis, critical in policing, defense and intelligence sectors, is already utilized as a means to monitor social media for potential threats (Gomez-Perez et al., 2022). These not only include criminal activities, but also accidents and natural disasters. Similarly, governments and political parties are also turning towards LT to analyse political discourse and public opinion, useful when appraising public feedback or predicting election outcomes.

2.7 The Media, Truth, Trust, and Accuracy in Reporting

Transformations in news media are also being driven by digital technology. On the one hand, digital platforms have disrupted traditional revenue streams, forcing traditional news outlets to restructure their business models. On the other, perpetual digital access to news content has made it necessary for news organizations to constantly update their online content. The pressure to provide the latest information has resulted in increased difficulty to verify truthfulness and accuracy in reporting. One result has been an erosion in trust, exacerbated by the rise of post-truth politics, the relativisation of facts, and the belief that personal perspectives carry more weight than objective reporting. In this respect, LT and AI may be able to help slow misinformation generated as a result of social media and post-truth politics. Political bias in journalism can also be detected and addressed through LT, a necessary tool during this era of growing polarisation and doubt over veracity.

2.8 Digital Technologies and Healthcare

The use of data and technology is influencing healthcare as well. E-health approaches, personalized medicine, digitalization, and large datasets are contributing to the way medicine is practiced today (E and E, 2020; S et al., 2020; ALLEA et al., 2021).[43] The trend towards hyperconnectivity may be seen in wearables, including health-oriented devices, that provide a

[42] https://www2.deloitte.com/xe/en/insights/industry/public-sector/government-trends/2021/digital-government-transformation-trends-covid-19.html

[43] https://knowledge4policy.ec.europa.eu/shifting-health-challenges_en;https://knowledge4policy.ec.europa.eu/foresight/digitize-me-my-health_en

means to measure or detect events around or inside one's body in real time. Similarly, digital technology and AI are aiding in the development of healthcare applications, including remote monitoring and AI-supported diagnostic devices. Greater investment is being made into virtual cognitive agents, such as medical billing assistants, radiology assistants, plan of care assistants, and medical testing assistants. Virtual assistants such as these can provide the public with access to trustworthy information and the market for virtual medical assistants is expected to grow significantly over the next few years. In addition, medical transcription tools, a growing area in the health domain, can help doctor-patient interactions. Automatic transcription not only standardises note taking, but also allows doctors to focus on patients without the need to produce notes manually. Through these digital technologies, prevention, diagnosis, treatment, and management of health-related issues can be improved and patients are afforded the opportunity to engage more flexibly with healthcare providers. Indeed, digital and AI technologies have begun to enable more personal control over health, sometimes allowing patients to care for themselves from home via virtual communication, a trend that is expected to grow and normalize, possibly with some required improvements in terms of awareness and effectiveness. Digital technologies can also aid in delivering critical health-related information to the public efficiently and effectively. As the COVID-19 pandemic illustrated, Machine Translation (MT) proved useful in translating health guidelines, recommendations and information across a wide range of languages (Bērziņš et al., 2022).

2.9 LT and Migration

Because migration patterns can be difficult to predict, it is unclear if Europe will experience a significant influx of immigrants over the next decade. However, current immigration and intra-European migration trends demonstrate that language differences can pose barriers to effective integration for newcomers. This is a problem that AI and LT can ameliorate. Healthcare represents just one domain in which LT may alleviate some of the many complications associated with language barriers that arise due to migration. One example is the problem of doctor-patient communication. MT can aid in constructing patient histories, communicating recommended treatments, and establishing direct dialog between doctor and patient (Bērziņš et al., 2022). The same may be said for other areas of healthcare, such as communication between patient and insurer.

2.10 Gaming and Entertainment

Entertainment is another area in which digital technologies and AI are making an indelible mark. One potential trend in which MT may play a greater role in the future, for instance, is video game localisation, a field that requires cross-language communication and knowledge of cultural contexts. Apart from a variety of texts that need translation within games and the gaming industry, it is increasingly apparent that a kind of digital "universal translator" would benefit online multiplayer games that feature in-game dialogue and collaboration from players across the globe. Moreover, given that games are generally translated into only a handful of the more dominant languages, this is yet another area where lesser-spoken languages could benefit from MT.

2.11 Possible Downsides and Guardrails

There are, nonetheless, downsides to the expanding trend in digitization. The widening impact of the digital and data-driven world touches upon a range of social and ethical questions in contemporary life. While the ability to collect and analyze massive amounts of data has enabled intriguing breakthroughs, it has also brought new threats that must be mitigated,

such as the potential for cyberattacks that seek to undermine public trust in democratic institutions and challenge the core values of societies. Increased reliance on digital technologies has opened the door to identity theft, disruptions to infrastructures, and the misuse of personal data, a concern that is especially marked with respect to control over healthcare data. And because the internal workings of current language models are not yet fully understood, AI can also produce misleading results in ethically challenging activities, such as political profiling, job screening, policing, and surveillance, generating fears over the erosion of civil liberties. In the economic arena, concern about the possible effects of AI and digitalization on the workforce will need to be addressed and appropriate measures taken to foster a smoother transition. Among the possible negative trends is the danger that AI may give executives and managers excessive control over employees. Ideally, a comprehensive policy approach including adequate regulation and investment could help avoid the pitfalls currently associated with AI. If handled prudently, AI and LT may strengthen European sovereignty and enhance well-being. For this to happen, European interests and values concerning data use must be asserted. Europe should set parameters around who accesses and utilizes European data, an idea that is bolstered by Europe's legal framework for data protection. AI's impact on Europeans' lives should be controlled by Europeans, who have an opportunity to establish a model that utilizes AI to benefit society. To do this, Europe should develop a coordinated AI strategy that is built upon local data-sharing ecosystems, which can aid in developing local solutions for safe and socially ethical AI development.

3 Language Technology and Language-Centric Artificial Intelligence

3.1 Language Technology: A Brief History and General Overview

Understanding language is key for building intelligent systems. In fact, most of the digital information available is unstructured information in the form of documents (written or spoken) in multiple languages, representing a challenge for any organization that wants to exploit and process its information. In fact, up to 80% of all data is unstructured text data.[44] Most computer systems process only structured data (for example databases with millions of records) because it is non-trivial to process *unstructured digital information* (including written and spoken language). Unstructured language data is subject to multiple interpretations (ambiguity), requires knowledge about the context and the world and it is intrinsically complex to process.

Interest in the computational processing of human languages (machine translation, dialogue systems, etc.) coincided with the emergence of AI and, due to its increasing importance, the discipline has been established as specialized fields known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or LT. While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In practice, these communities work closely together, sharing the same publishing venues and conferences, combining methods and approaches inspired by both, and together making up *language-centric AI*. In this report we treat them interchangeably as long as it is not otherwise explicitly stated.

LT is concerned with studying and developing systems capable of processing human language. The field has developed, over the years, different methods to make the information contained in written and spoken language explicit or to generate or synthesise written or spoken language. Despite the inherent difficulty of many of the tasks performed, current LT support allows many advanced applications which have been unthinkable only a few years

[44] http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

ago. LT is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, text editors, text predictors, automatic translation systems, automatic subtitling, automatic summaries, inclusive technology, etc. Its rapid development in recent years predicts even more encouraging and also exciting results in the near future

LT has come far in the nearly three quarters of a century since its beginnings as a discipline in the 1950s, when Alan Turing outlined his famous criterion to determine whether a machine could be considered intelligent (Turing, 1950). Not long after, Noam Chomsky laid the foundations to formalise, specify and automate linguistic rules with his generative grammar (Chomsky, 1957). The horizon set by Turing and the instrument provided by Chomsky influenced the vast majority of NLP research for years to come. This early era in LT was closely linked to MT in the belief that a quality automatic translator would soon be in hand. By the mid-1960s, however, the Automatic Language Processing Advisory Committee (ALPAC) report, issued by a panel of leading US experts acting in an advisory capacity to the US government, revealed the true difficulty of the task and NLP in general (Pierce and Carroll, 1966). The ALPAC report had a devastating impact on R&D&I funding for the field and the NLP community turned towards more realistic objectives.

The following two decades were heavily influenced by Chomsky's ideas, but by the late 1980s the seeds of a revolution that would irreversibly alter NLP were planted. This upheaval was driven by four factors: 1) the clear definition of individual NLP tasks and corresponding rigorous evaluation methods; 2) the availability of relatively large amounts of data; 3) computers that could process these large amounts of data; and 4) the gradual introduction of more robust approaches based on statistical methods and Machine Learning (ML). As the new millennium neared and unfolded, these elements paved the way for major subsequent developments. In addition to a host of novel tools and applications, several wide-coverage linguistic resources, such as WordNet (Miller, 1992), were created that reshaped the field. Data-based systems began to displace rule-based systems, leading to the almost ubiquitous presence of ML-based components in NLP systems. Collobert et al. (2011) presented a multilayer neural network adjusted by backpropagation that solved various sequential labeling problems. Word embeddings gained particular relevance due to their role in allowing the incorporation of pretrained external *knowledge* into neural architecture (Mikolov et al., 2013b; Pennington et al., 2014; Mikolov et al., 2018). Large volumes of unannotated texts, together with progress in self-supervised ML and the rise of high-performance hardware in the form of Graphic Processing Units (GPUs), enabled highly effective deep learning systems to be developed across a range of application areas. These and other breakthroughs characterized the radical technological shift that took place in NLP in the 2010s and helped launch today's Deep Learning Era.

3.2 State of the Art

Around ten years ago, *Deep Learning* (Salakhutdinov, 2014) started gaining traction in LT thanks to mature deep neural network technology, much larger datasets, more computational capacity (notably, the availability of GPUs), and application of simple but effective self-learning objectives (Goodfellow et al., 2016). One of the advantages of these neural language models is their ability to alleviate the *feature engineering* problem by using low-dimensional and dense vectors (aka. *distributed representation*) to implicitly represent the language examples (Collobert et al., 2011). By the end of 2018,[45] the field of NLP observed another relevant disruption with BERT (Devlin et al., 2019). Since then BERT has become a ubiquitous baseline in NLP experiments and inspired a large number of studies and improvements (Rogers et al., 2020). This pretrained language model recipe has been replicated across

[45] The paper first appeared in http://arxiv.org.

languages leading to many language specific BERTs such as FlauBERT and CamemBERT for French (Le et al., 2020; Martin et al., 2020), Robbert for Dutch (Delobelle et al., 2020), BERTeus for Basque (Agerri et al., 2020), etc.

LT is undergoing a paradigm shift with the rise of *neural language models*[46] that are trained on broad data at scale and are adaptable to a wide range of monolingual and multilingual downstream tasks (Devlin et al., 2019; Qiu et al., 2020; Liu et al., 2020; Torfi et al., 2020; Wolf et al., 2020; Han et al., 2021; Xue et al., 2021). Though these models are based on standard *self-supervised* deep learning and *transfer learning*, their scale results in new emergent and surprising capabilities.

In *self-supervised learning*, language models are derived automatically from large volumes of unannotated language data (text or speech). There has been considerable progress in *self-supervised learning* since *word embeddings* (Turian et al., 2010; Mikolov et al., 2013a; Pennington et al., 2014; Mikolov et al., 2018) associated word vectors with context-independent vectors. Shortly thereafter, self-supervised learning based on autoregressive language modelling (predict the next word given the previous words) (Dai and Le, 2015) became popular. This approach produced language models such as GPT (Radford et al., 2018), ELMo (Peters et al., 2018) and ULMFiT (Howard and Ruder, 2018). The next wave of developments in self-supervised learning — BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), BART (Lewis et al., 2020) — quickly followed, embracing the Transformer architecture (Vaswani et al., 2017), incorporating more powerful deep bidirectional encoders of sentences, and scaling up to larger models and datasets. Open-source libraries such as Transformers[47] may open up these advances to a wider LT community. The library consists of carefully engineered state-of-the art Transformer architectures under a unified API and a curated collection of pretrained models (Wolf et al., 2020). For example, BERT (Devlin et al., 2019) applies two training self-supervised tasks namely *Masked Language Model* and *Next Sentence Prediction*. The *Masked Language Model* learns to predict a missing word in a sentence given its surrounding context while the *Next Sentence Prediction* learns to predict if the next sentence will follow the current one or not. Self-supervised tasks are not only more scalable, just depending on unlabelled data, but they are designed to force the model to predict coherent parts of the input. Through self-supervised learning, tremendous amounts of unlabeled textual data can be utilised to capture versatile linguistic knowledge without labour-intensive workloads.

The idea of *transfer learning* is to take the "knowledge" learned from one task (e.g., predict the next word given the previous words) and apply it to another task (e.g., summarization). With transfer learning, instead of starting the learning process from scratch, you start from patterns that have been learned when solving a different problem. This way you leverage previous learning and avoid starting from scratch. Within deep learning, pretraining is the dominant approach to *transfer learning*: the objective is to *pretrain* a deep transformer model on large amounts of data and then reuse this pretrained language model by *fine-tuning* it on small amounts of (usually annotated) task-specific data. Thus, transfer learning formalises a two-phase learning framework: a pretraining phase to capture knowledge from one or more source tasks, and a fine-tuning stage to transfer the captured knowledge to many target tasks.

Recent work has shown that pretrained language models can robustly perform classification tasks in a few-shot or even in zero-shot fashion, when given an adequate task description in its natural language prompt (Brown et al., 2020; Ding et al., 2021). Unlike traditional supervised learning, which trains a model to take in an input and predict an output, *prompt-based learning* is based on exploiting pretrained language models to solve a task using text directly (Liu et al., 2021). To use these models to perform prediction tasks, the original input

[46] Also known as Pretrained Language Models (Han et al., 2021)
[47] https://huggingface.co/

is modified using a template into a textual string prompt that has some missing slots, and then the language model is used to probabilistically fill the missing information to obtain a final string, from which the final output for the task can be derived. This framework looks very promising for a number of reasons: it allows the language model to be pretrained on massive amounts of raw text, and by defining a new prompting function the model is able to perform few-shot or even zero-shot learning, adapting to new scenarios with few or no labeled data. Thus, some NLP tasks can be solved in a fully unsupervised fashion by providing a pretrained language model with "task descriptions" in natural language (Raffel et al., 2020; Schick and Schütze, 2021). Surprisingly, fine-tuning pretrained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Wei et al., 2021; Sanh et al., 2021; Min et al., 2021; Ye et al., 2021; Aghajanyan et al., 2021; Aribandi et al., 2021).

In some cases, an increase in scale has led to such behavior. One of the largest dense language models, GPT-3 (Brown et al., 2020), for instance, is able to perform tasks that it was not explicitly trained to solve with zero to few training examples (referred to as zero-shot and few-shot learning, respectively).[48] Not only was this ability mostly absent from its predecessor GPT-2, over 100 times smaller than GPT-3, but the latter also outperforms state-of-the-art models on certain tasks for which they *were* explicitly trained to solve. It is impressive that models such as GPT-3 can achieve state-of-the-art performance in limited training data regimes. Most models developed until now have been designed for a single task and thus can be evaluated effectively by a single metric. The eye-opening results has encouraged various IT enterprises, including Google, Microsoft and OpenAI, to develop and deploy their own large pretrained neural language models. Fortunately, there are also open source alternatives to GPT-3. For instance, GPT-Neox-20b is a 20 billion parameter autoregressive language model trained on the Pile (Black et al., 2022) and OPT is a series of open-sourced large causal language models which perform similar in performance to GPT-3 Zhang et al. (2022).

Multilingual Language Models (MLLMs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), mBART (Liu et al., 2020), BLOOM,[49] etc. have emerged as a viable option for bringing the power of pretraining to a large number of languages. An MLLM is pretrained using large amounts of unlabeled data from multiple languages with the hope that low-resource languages may benefit from high-resource languages due to a shared vocabulary and latent language properties. For example, mBERT (Devlin et al., 2019) pretrained using non-parallel multilingual Wikipedia corpora in 104 languages, has the ability to generalize across languages in zero-shot scenarios. This indicates that even with the same structure of BERT, using multilingual data can enable the model to learn cross-lingual representations. The surprisingly good performance of MLLMs in crosslingual transfer as well as bilingual tasks motivates the hypothesis that MLLMs are learning universal patterns (Doddapaneni et al., 2021). Thus, one of the main motivations of training MLLMs is to enable transfer from high resource languages to low-resource languages. Thus, of particular interest is the ability of MLLMs to facilitate zero-shot crosslingual transfer from a resource-rich language to a resource-deprived language which does not have any task-specific training data, or to fine-tune more robust language models by using annotated training data in multiple languages.

The BigScience[50] community-based initiative has recently released BLOOM[51], the first multilingual large language model trained in complete transparency. BLOOM is the result of the largest collaboration of AI researchers ever involved in a single research project Le Scao

[48] GPT-3 can be fine-tuned for an excellent performance on specific, narrow tasks with very few examples. It possesses 175 billion parameters and was trained on 570 gigabytes of text, with a cost estimated at more than four million USD (https://lambdalabs.com/blog/demystifying-gpt-3/).
[49] https://bigscience.huggingface.co/blog/bloom
[50] https://bigscience.huggingface.co/
[51] https://bigscience.huggingface.co/blog/bloom

et al. (2022). With its 176 billion parameters, BLOOM is able to generate text in 46 natural languages and 13 programming languages. For almost all of them, such as Spanish, French and Arabic, BLOOM is be the first language model with over 100B parameters ever created. This is the culmination of a year of work involving over 1000 researchers from 70+ countries and 250+ institutions, leading to a final run of 117 days (March 11 - July 6) training the BLOOM model on the Jean Zay supercomputer[52] in the south of Paris, France thanks to a compute grant worth an estimated €3M from French research agencies CNRS and GENCI.

Despite their notable capabilities, however, large pretrained language models such as GPT-3 come with important drawbacks that will require interdisciplinary collaboration and research to resolve.[53] To begin with, we currently have no clear understanding of how they work, when they fail, and what emergent properties they present. Indeed, some authors call these models *foundation models* to underscore their critically central yet incomplete character (Bommasani et al., 2021). And because their defects are inherited by all adapted models downstream, their effectiveness across so many tasks demands caution. Second, the systems are extremely sensitive to phrasing and typos, are not robust enough, and perform inconsistently (Ribeiro et al., 2018, 2019). Additionally, existing laboratory benchmarks and datasets have numerous inherent problems; the ten most cited AI datasets are riddled with label errors, which are likely to distort our understanding of the field's progress (Caswell et al., 2021; Northcutt et al., 2021). Third, these models are expensive to train, which means that only a limited number of organisations can currently afford to construct such models. There is a growing concern that this is fostering unequal access to computing power, providing undue advantages in modern AI research (Ahmed and Wahed, 2020) to determined companies and elite universities which possess abundant funding, computing capabilities, LT experts and data. Fourth, large NLP datasets, including one utilized to train Google's Switch Transformer and T5 model, can generate racist, sexist, and otherwise biased text when they are "filtered" to remove Black and Hispanic authors, material related to LGBTQ identities, and source data that deals with a number of other minorities (Dodge et al., 2021).[54] Moreover, large language models can sometimes produce unpredictable and factually inaccurate text or even recreate private information.[55] Finally, computing large pretrained models comes with a substantial carbon footprint.[56] Strubell et al. (2019b) recently estimated that the training process for one sizable neural architecture emitted 284 tons of carbon dioxide, almost 57 times the estimated amount that the average human is responsible for in a year.[57] In short, notwithstanding claims of human parity in many LT tasks or hype regarding machines being sentient,[58] Natural Language Understanding (NLU) is still an *open research problem* far from being solved since all current approaches have *severe* limitations.

3.3 Main Challenges

The current acceleration in AI and LT will have a fundamental impact on society, as these technologies are at the core of the tools we use on a daily basis.

[52] http://www.idris.fr/eng/jean-zay/cpu/jean-zay-cpu-hw-eng.html
[53] https://lastweekin.ai/p/the-inherent-limitations-of-gpt-3
[54] https://www.unite.ai/minority-voices-filtered-out-of-google-natural-language-processing-models/
[55] https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html
[56] https://spectrum.ieee.org/deep-learning-computational-cost
[57] https://ourworldindata.org/co2-emissions
[58] https://www.washingtonpost.com/business/do-computers-have-feelings-dont-let-google-alone-decide/2022/06/14/0e6c0d3a-ebaf-11ec-9f90-79df1fb28296_story.html

3.3.1 Language Models and Language Diversity

Recent progress in LT has been driven by advances in both deep learning model architectures and large neural model pretraining. Transformer architectures have facilitated the building of higher-capacity models and pretraining has made it possible to effectively utilise this capacity for a wide variety of languages and tasks. Unfortunately, the resources necessary to create the best-performing neural language models are developed almost exclusively by US and China technology giants. Moreover, this transformative technology poses problems from a research advancement, environmental, and ethical perspective. For example, models such as GPT-3 are private, anglo-centric, and inaccessible to academic organisations (Floridi and Chiriatti, 2020; Dale, 2021). This situation also promotes a colossal duplication of energy requirements and environmental costs, due to the duplicated training of private models. In addition, there are worrying shortcomings in the text corpora used to train these models, ranging from a lack of representation of populations, to a predominance of harmful stereotypes, and to the inclusion of personal information.

Given these issues and the role of LT in everyone's daily lives, many LT practitioners are particularly concerned by the lack of language diversity in LT research and the need of transparent digital language equality across all aspects of European society, from government to business to citizens.[59] Looking ahead, it is possible to foresee intriguing opportunities and new capabilities in this regard, but also a range of uncertainties and inequalities that may leave several groups disadvantaged Sayers et al. (2021). Joshi et al. (2020), for instance, examine the relationship between types of languages, resources and their representation in conferences over time. As expected, only a small number of world's over 7000 languages are represented in the rapid evolving LT field. This disproportionate representation if further exacerbated by systematic inequalities in LT across the world's languages. After English, only a handful of Western European languages – principally German, French and Spanish – and even fewer non-Indo-European languages – primarily Chinese, Japanese and Arabic – dominate the field. Blasi et al. (2021) suggest that this is because LT development is driven by the economic status of the language users, rather than the sheer demographic demand. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pretrained language models, prompt learning and self-supervised systems opens up the way to leverage LT for less developed languages.[60] For the first time, a single multilingual model has outperformed the best specially trained bilingual models on news translations. That is, a single multilingual model provided the best translations for both low- and high-resource languages, showing that the multilingual approach is indeed the future of MT (Tran et al., 2021). However, the development of these new LT systems would not be possible without sufficient resources (experts, data, computing facilities, etc.) along with carefully designed evaluation benchmarks and annotated datasets for every language and domain of application.

Forecasting the future of LT and language-centric AI is a challenge. A decade ago, few would have predicted the recent breakthroughs that have resulted in systems that translate without parallel corpora (Artetxe et al., 2019), create image captions (Hossain et al., 2019), generate pictures from textual descriptions (Ramesh et al., 2021),[61] produce playscripts (Rosa et al., 2020), yield text that is nearly indistinguishable from human prose (Brown et al., 2020), provide high quality explanations for novel jokes not found on the web (Chowdhery et al., 2022) and successfully solve unseen tasks (Wei et al., 2021; Sanh et al., 2021; Min et al., 2021; Ye et al., 2021; Aghajanyan et al., 2021; Aribandi et al., 2021). Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pretrained language models and self-supervised systems opens up the way to leverage LT for less developed languages. Furthermore, inspired by progress in large-scale language modeling, similar ap-

[59] https://gitlab.com/ceramisch/eacl21diversity/-/wikis/EACL-2021-language-diversity-panel
[60] https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html
[61] https://openai.com/blog/dall-e/

proaches towards building a single generalist agent are being developed beyond the realm of text outputs. For instance, the agent called Gato works as a multi-modal, multi-task, multi-embodiment generalist policy. Gato can perform over 600 various tasks, including play video games, caption photos, and move real-world robotic arms.[62] It is, nevertheless, safe to assume that many more advances will be achieved utilizing pretrained language models and that they will impact society unpredictably. Future users are likely to discover novel applications and wield them positively (such as knowledge acquisition from electronic health records) or negatively (such as generating deep fakes). In either case, as argued by Bender et al. (2021), it is important to understand the current limitations of large pretrained language models, which they call "stochastic parrots," and put their successes in context. Focusing on state-of-the-art results exclusively with the help of leaderboards, without encouraging deeper understanding of the mechanisms by which they are attained, can give rise to misleading conclusions. These, in turn, may direct resources away from efforts that would facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication.

3.3.2 Natural Language Understanding

Despite recent progress in AI and NLP producing algorithms that perform well on a number of LT tasks, it is still unclear how to move forward and develop algorithms that understand language as well as humans do. Many limitations of today state-of-the-art methods become evident when comparing with the human ability to understand language Gardner et al. (2019); Lake et al. (2017); Tamari et al. (2020); Bender and Koller (2020); Linzen (2020). Many cognitive scientists posit that humans create rich mental models of the world from their observations which provide superior explainability, reasoning, and generalizability to new domains and tasks Saparov and Mitchell (2022). How do we, as a field, move from today's state-of-the-art to more general intelligence? What are the next steps to develop algorithms that can generalize to new tasks at the same level as humans?

LT is a diverse field, and progress throughout its development towards NLU has come from new representational theories, modeling techniques, modalities, data collection paradigms, competitions and tasks. The present success of representation learning approaches trained on large, text-only corpora requires the parallel tradition of research on grounding the broader physical and social context of language to address the deeper questions of communication, commonsense and reasoning Bisk et al. (2020).

3.3.3 Data Resources and Benchmarking

Current LT research requires large coordinated and collaborative efforts with sufficient resources (experts, data, computing facilities, etc.) involving pan-european LT research centers, national and also regional administrations. For instance, the paper described the development of PaLM (Chowdhery et al., 2022) has been signed by 68 authors. Ambitious LT research can only be achieved by gathering the necessary resources in terms of data, computing facilities, expertise, etc. which is available by a small number of research labs in Europe. Virtual research laboratories joining efforts from small research labs from academia, nonprofit organizations or small companies can also gather the necessary critical mass, resources and interdisciplinary expertise to establish a coordinated world-class research collaboration in LT.

The real source of the most recent progress is the increase in data size and diversity. Models are only a reflection of their training data. Thus, access to sufficient multilingual and

[62] https://www.deepmind.com/publications/a-generalist-agent

multi-modal data of quality (responsible, legal, diverse, unbiased, ethical, representativeness, etc.), in all European languages and domains (media, health, legal, education, etc.) is one of the major challenges for developing the full potential of LT. Unfortunately, most of this data is currently inaccessible to European LT researchers.

Current LT research also requires flexible access to High Performance Computing (HPC) facilities in the form of clusters of high capacity GPUs. There are many EU initiatives ofering HPC: EuroHPC JU,[63] PRACE,[64] national computing facilities, etc. However, it is unclear if these initiatives are ready to provide the computing support that the European LT research community currently needs for developing state-of-the-art language models for all languages, domains, tasks and modalities. For instance, training the multilingual language model BLOOM on the Jean Zay supercomputer[65] took approximately one million compute hours on 117 days.[66] The hardware consists of the following:

- 384 NVIDIA A100 80GB GPUs (48 nodes) with 32 spare GPUs
- 8 GPUs per node, connected using NVLink 4, OmniPath
- Each node is powered by an AMD EPYC 7543 32-Core Processor, with a total of 512GB CPU memory and 640GB GPU memory

Access protocols are also different across computing facilities. Flexible access is needed for small experiments.[67] Much larger experiments for developing large language models on hundreds of GPUs should follow a more elaborated protocol. These HPC facilities should also provide flexible access to the LT industry. These HPC facilities should also provide clear and robust protocols to process sensible data.

Furthermore, assessing the real progress of LT also requires developing better benchmarks and datasets (ethical, responsible, legal, etc.) for all languages, domains, tasks and modalities.

Current leading research on LT is already multilingual covering hundreds of languages simultaneously and multi-modal including text, image, audio, video, interactions, etc. In fact, some aspects of world knowledge are difficult or impossible to learn from text only. This requires to extend the research limits of LT beyond language.

4 Language Technology and Digital Language Equality in 2022

4.1 Digital Language Equality in Europe: Where Are We Now?

As shown in the previous section, the LT field as a whole has shown remarkable progress during the last few years, especially in Europe, thanks in particular to substantial funding coming from the EU through various schemes. The advent of deep learning and neural networks over the past decade, together with the considerable increase in the number and quality of resources for many languages, has yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? In other words, do all European languages benefit equally and fairly from this overall progress, can they be considered digitally equal, in the interest of their speaker communities?

[63] https://eurohpc-ju.europa.eu/index_en
[64] https://prace-ri.eu/
[65] http://www.idris.fr/eng/jean-zay/cpu/jean-zay-cpu-hw-eng.html
[66] https://bigscience.huggingface.co/blog/bloom
[67] https://www.edari.fr/schema/acces/ressource

To answer these questions, the following definition of Digital Language Equality (DLE) has been introduced and adopted: **Digital Language Equality** is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age (Gaspari et al., 2022b).

This definition provides the basis to establish a metric that enables the quantification of the level of technological support for each language in scope of ELE with descriptive, diagnostic and predictive value to successfully promote DLE. This approach facilitates comparisons across languages, tracking their advancement towards the goal of DLE, as well as the prioritisation of needs, especially to fill existing gaps, focusing on realistic and feasible targets. The **DLE Metric** is therefore defined as "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" (Gaspari et al., 2021). The DLE Metric is computed for each language on the basis of various factors, grouped into **technological factors** (or TFs, e. g. the available language resources, tools and services, which are the focus of this paper) and situational **contextual factors** (or CFs, e. g. societal, economic, educational, industrial, etc. conditions).

More specifically, the TFs are divided into two main categories, that are further broken down into more specific sub-categories. The first one includes tools and services that are offered via the web or running in the cloud, but also downloadable tools, source code, etc.; this category encompasses, for example, NLP tools (morphological analysers, part-of-speech taggers, lemmatisers, parsers, etc.); authoring tools (e. g. spelling, grammar and style checkers); services for information retrieval, extraction, and mining, text and speech analytics, MT, natural language understanding and generation, speech technologies, conversational systems, etc.

The second category of TFs includes datasets, i. e. corpora or collections of text documents, text segments, audio transcripts, audio and video recordings, etc., monolingual or bi- /multilingual, raw or annotated. It also encompasses language models and computational grammars and lexical and conceptual resources, including 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.), such as computational lexica, gazetteers, ontologies, term lists, thesauri, etc.

To objectively and consistently quantify the TFs for all of Europe's languages, we assigned weights for the computation of the DLE Metric formula, as described in detail in Gaspari et al. (2022a). The weights are assigned to the features and relevant values that are recorded in the European Language Grid (ELG) Catalogue, where resources, be they datasets or tools, are entered with rich and fine-grained accompanying metadata. The resulting technological DLE score is open-ended in principle, i.e. it increases for a language as new datasets and tools that are relevant to that language are added to the ELG Catalogue, alongside the respective metadata. The DLE Metric can be computed dynamically and the scores can be interactively visualised in real time on the basis of the data available in the ELG Catalogue via the ELE/ELG Dashboard.[68]

Figure 1 shows the technological DLE scores for all the languages covered by ELE as of 30th June 2022, based on the data visualisation offered by the ELE/ELG Dashboard. English clearly leads the way with a technological DLE score of 63,893, that is nearly twice as much that of the following two languages, namely German (34,196) and Spanish (32,301). French follows in fourth position with a much lower score of 26,750, and then there is a further significant drop in scores, even for languages with large populations of speakers that are official EU languages. Overall, the official EU and national languages are clustered towards the left of Figure 1, many of them with very modest scores compared to English, German and Spanish. For example, Greek and Czech, with technological DLE scores of 9,951 and 9,790,

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

Figure 1: Technological DLE scores as of 17th October 2022

respectively, are in 11th and 12th position, but their scores are less than a sixth of that for English. The vast majority of the languages, starting from Norwegian, Basque and Turkish, currently have technological DLE scores under 5,000, and well over half of the languages included in Figure 1, especially minority and regional ones, have technological DLE scores scores in three digits or less; in several cases the scores are very close to zero. This situation shows a very clear and alarming imbalance in the current LT support for Europe's languages, that needs to be addressed promptly to move in the direction of DLE in Europe by 2030.

In addition to the TFs, the DLE metric also includes the CFs, that represent the "general conditions and situations of the broader context" of the language communities (Gaspari et al., 2021). A language with a high contextual DLE score enjoys a context with the possibility to evolve, supported by political will, potential for funding, innovation and economic interest, while a language with a low score finds itself in the opposite position. Therefore, the score calculated for the contextual factors indicates the potential of the language to achieve DLE. The specific elements that make up the CFs and the formula to compute them are described in detail in Grützner-Zahn and Rehm (2022). The contextual DLE score for each language covered by ELE is the result of an averaging process, where the score for each language is relative to the others, therefore the relevant values range between 0 and 1.

Figure 2 shows the contextual DLE scores for all the languages covered by ELE as of 30th June 2022, as provided by the ELE/ELG Dashboard. The contextual DLE scores show a strong tendency to provide a high score for the official EU languages with the largest language communities, and a low score for the regional and minority languages of Europe. Official national languages which are not also official EU languages rank in between. In order to achieve DLE in Europe, the languages with a low score must be specifically supported, e.g. for the development of high-quality LRTs through dedicated funding. For these languages, there is no context that would otherwise enable the development of LRTs. Given that at present many languages with the lowest contextual scores have few or even no LRTs in the ELG Catalogue, these languages face a real danger of digital extinction, because very little or no digital support is available at the moment for them and the conditions do not indicate a promising situation.

A key 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 check the progress or the status of one or more European languages with respect to the overall goal of achieving DLE. In particular, with regard to the TFs, as the ELG Catalogue organically grows over time, the resulting technological DLE 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 has available, also showing where the status is less than ideal or not at the expected level. Similarly, the contextual DLE scores can be updated when the high-quality relevant sources that have been used to compute them release updated data.

While the technological and contextual scores of the DLE metric are illustrative of the overall situation for Europe's languages, a more detailed investigation of the LRTs currently available per language on the Catalogue of the ELG platform allows us to more accurately identify gaps and inequalities in particular LT application areas and for specific basic language resource types. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into a level of technology support.

As expected, the best supported language is English, the only language that is classified in the *good support* group. English is primarily acting as a benchmark for the level of technological support that other European languages could receive. While it is extremely unlikely that any other European language will reach this level, due to the continuing development of support for English, and thus serves as a moving goalpost, nevertheless it provides a good criterion for relative assessment. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g. Ger-

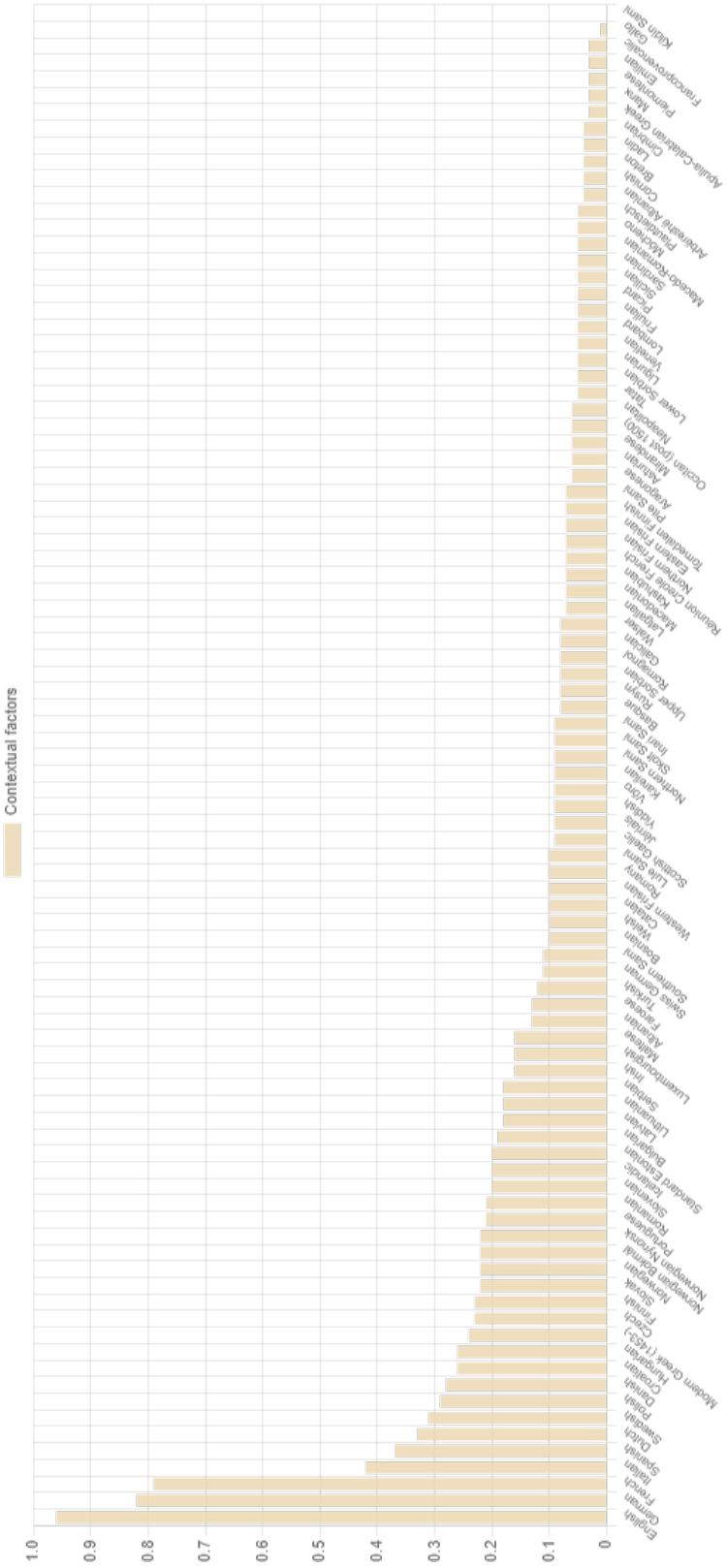

Figure 2: Contextual DLE scores as of 17th October 2022

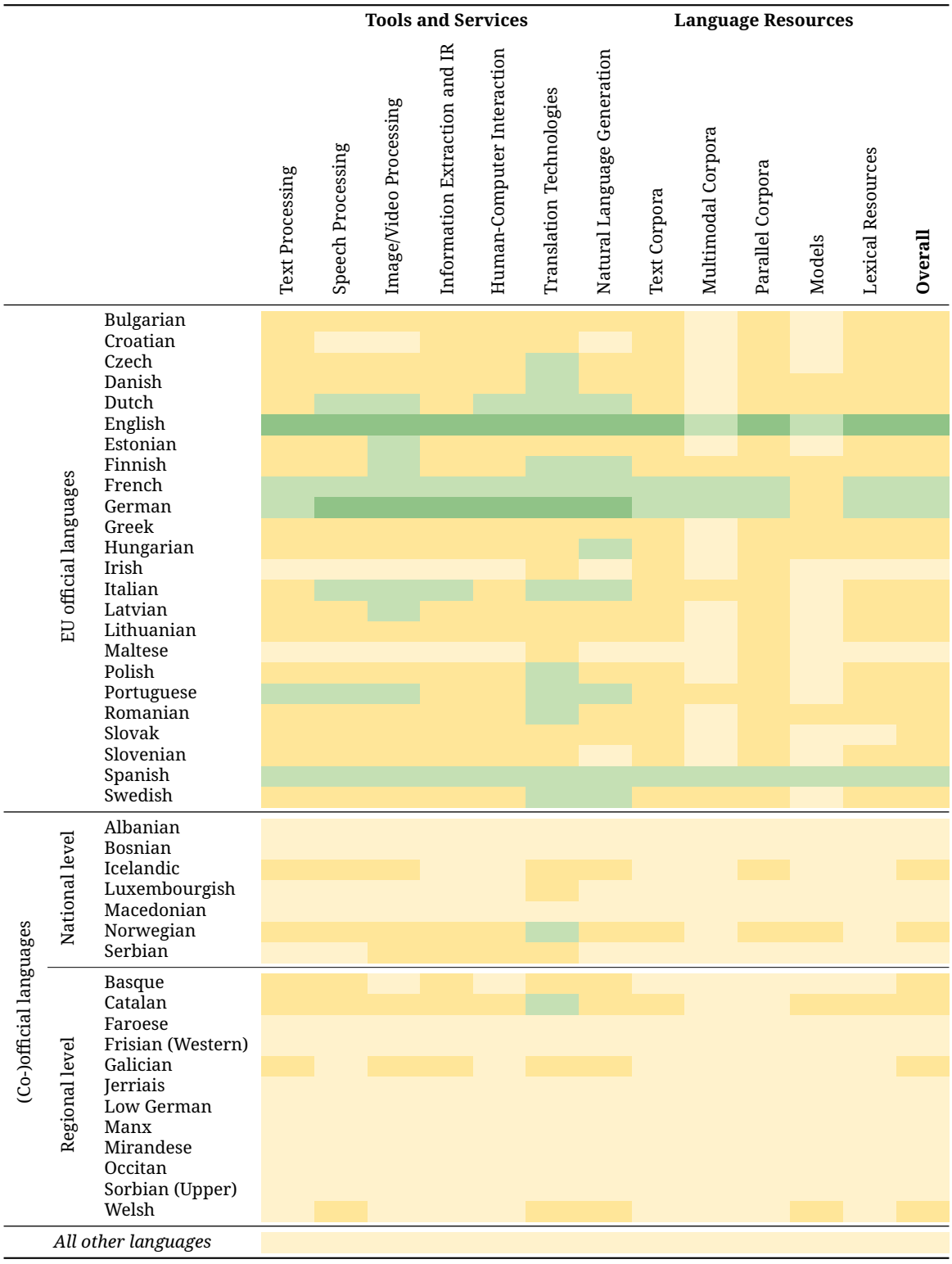

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

man in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[69] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases: while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all.

While a fifth level, *excellent support*, could have been foreseen, the situation is such that at present none of Europe's languages qualifies for this ambitious status. Currently no natural language is optimally supported by technology, i. e. the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural Language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and imbalance among the languages, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). In the 2012 study technology support for a total of 31 European languages was investigated and each language was assigned to one of five categories (from *weak/no support* to *excellent support*) with regard to four broad areas of the LT field: speech processing, machine translation, text analysis and speech and text resources. In the 2012 analysis a language was considered "in danger of digital extinction" if it was classified as having "weak/no support" in at least *one* of the four categories. Using this definition, it was found that as many as 22 of the 31 languages in total were in danger of digital language extinction.

The present analysis based on data from 2022 differs from the 2012 analysis in various respects, one of which is the number of categories we use to assess the technology support of a language. We now use 12 categories (instead of four), grouped into *tools and services* and *language resources* (Table 1). As this setup is more complex and more fine-grained, we have modified our definition of digital language extinction accordingly: a language is now considered "in danger of digital extinction" if it is classified into "weak/no support" in at least *two* of the 12 categories (Table 1).

Table 2 indicates that back in 2012, a total of 22 of the 31 languages under investigation (71.0%) were in danger of digital extinction. This number has risen to 75 out of 88 languages under investigation (85.2%) in 2022, which is due to the simple fact that we now take into account not only many more languages but many more languages with small or very small numbers of speakers.

In contrast, in 2012, a total of 16 official European Union languages were considered to be in danger of digital extinction. In 2022, this number has reduced to 11 languages.[70] At

[69] In addition to the languages listed in Table 1, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

[70] The official EU languages still in danger of digital extinction are Bulgarian, Croatian, Czech, Estonian, Irish, Latvian, Lithuanian, Maltese, Polish, Slovak, Slovenian.

Year	Project	Subset of Languages	Languages analysed	Languages in danger of digital extinction	Perc.
2012	META-NET	All META-NET languages	31	22	71.0%
		All EU languages	24	16	66.6%
		Co-official (national)	3	3	100%
		Co-official (regional)	4	3	75.0%
2022	ELE	All ELE languages	88	75	85.2%
		All EU languages	24	11	45.8%
		Co-official (national)	7	7	100%
		Co-official (regional)	12	12	100%
		All other languages	45	45	100%
		All META-NET languages	31	18	58.1%

Table 2: European languages in danger of digital extinction – 2022 vs. 2012

the same time, if we now compare the technology support of the 31 languages that were examined in 2012 with the current situation in 2022, we find that, all in all, 18 European languages are still in danger of digital extinction (compared to 22 in 2012). While the above-mentioned five official EU languages appear to be no longer threatened, due to the more fine-grained analysis scheme we apply in our current study, we now have to consider one language, Catalan, in danger of digital extinction in 2022, while it appeared to be not in danger in 2012. Catalan joins the group of co-official languages at the national or regional level, all of which are now considered in danger of digital extinction.

The complexities of the analyses clearly differ across the 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e.g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally supported ones in Europe is still evident in 2022. It is exactly this distance that needs to be ideally eliminated, or at least reduced, in order to move towards Digital Language Equality and avert the ever present risks of digital language extinction.

4.2 Europe's Languages in the Digital Sphere: Demands and Issues

As argued in Section 4.1, acute digital inequality exists between European languages. Moreover, the striking asymmetry between official and non-official EU languages with respect to available digital resources is worrisome. But even across the EU, there is an uneven distribution of resources (funding, open data, language resources, scientists, experts, computing facilities, IT companies, etc.) by country, region and language. In addition to the support of

a language per se, language varieties, dialects or accents may not be supported or only supported on very limited levels. At the same time, technological support is not a binary variable; it is rather a continuous variable with different shades and levels, as the technological and contextual DLE scores also indicate. This means that even in cases when a technology is available for a number of different languages, performance and accuracy typically vary across languages. In extreme cases, selected functionalities and/or support for minor languages may not be available at all. LTs are thus not accessible nor available to everyone on an equal level, i. e., functions, performance, robustness may be dramatically different from case to case.

In the following subsections we systematically present the issues pertinent to digital inequality of European languages and the recommendations of the ELE language informants in four axes: Data, Technologies, Compute and Research Infrastructures and Situational context.

4.2.1 Data

As evidenced, an abundance of training data for developing LTs is available only for a few languages with high commercial interest. For many (the majority of) European languages, this is not the case and only corpora which are minuscule in comparison to English are available. Since the development of LT systems is not possible without sufficient resources, the continuous collection and annotation of data as well as the creation of carefully designed and constructed evaluation benchmarks for every language and domain of application are some of the most prominent recurring issues and demands. This does not mean that each and every language needs to reinvent the wheel. It has been emphasised that the local LT communities should build on the previous results and best practices, to sustain, improve, consolidate and further develop existing tools and data resources.

The types of data needed for each language are of course different and their prioritisation should be based on the identification of gaps per language. For this reason it has been suggested that each country/region needs to define a strategic roadmap for identifying, building, curating, annotating and securing resources for varieties or domains that are critical for the local research, industry or for the administration. Language resource creation and compilation should be supported at national/regional level.

Although the data gaps per language are different, some data types have been frequently mentioned as priorities for many languages. These include: massive language models, both monolingual and multilingual; multimodal data, especially speech in conversational settings (dialogues) from speakers of different ages, genders and linguistic/dialectal backgrounds, but also video corpora for Sign Languages; domain-specific data (e. g. medical, legal or media among many others of interest); data for language use on social media; semantic resources (e. g. semantic annotations and knowledge bases); data for language pathologies; benchmarks, i. e. well-designed gold-standard corpora for fine-tuning language models and evaluating LT systems.

When investigating the current availability of some of the data types mentioned in the previous paragraph, as represented in the resources hosted in ELG on July 2022, it is apparent that even the best supported languages in this dimension, Spanish and English, are still only moderately covered (Figure 3). With respect to multimodal data, all languages with the exception of English, are weakly covered, with some, e. g. Maltese and Luxembourgish, severely underrepresented (Figure 4).

Apart from the “traditional” instruments for data collection, alternative approaches have been put forward by the ELE informants, such as automatic data generation and translation from more resourced languages. This can be for instance the case of many low-resource languages, such as regional languages, or languages that have historically co-existed with other

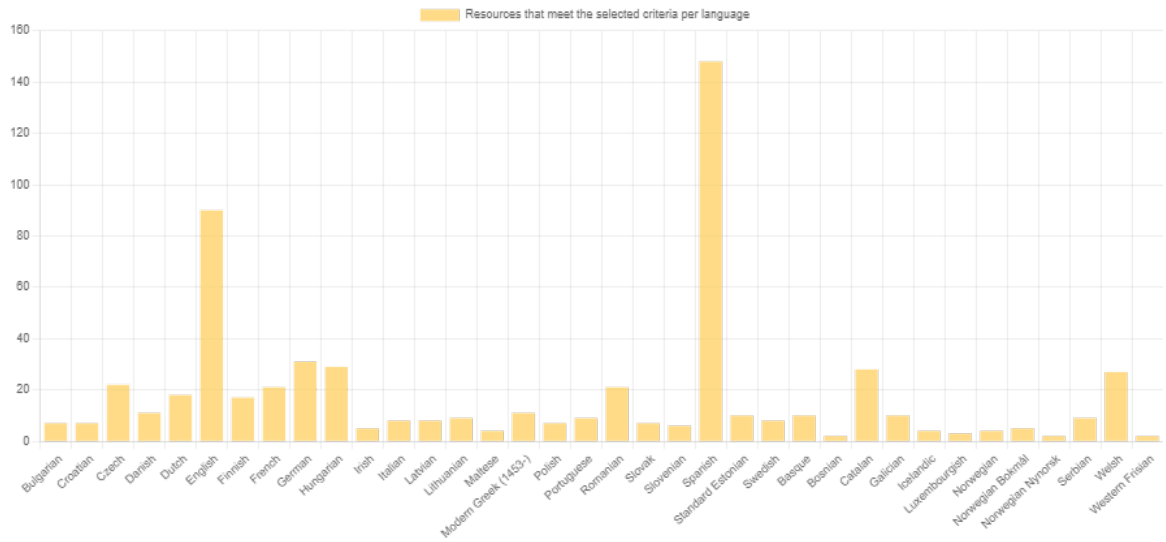

Figure 3: Number of language models available at the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones on 17.10.2022

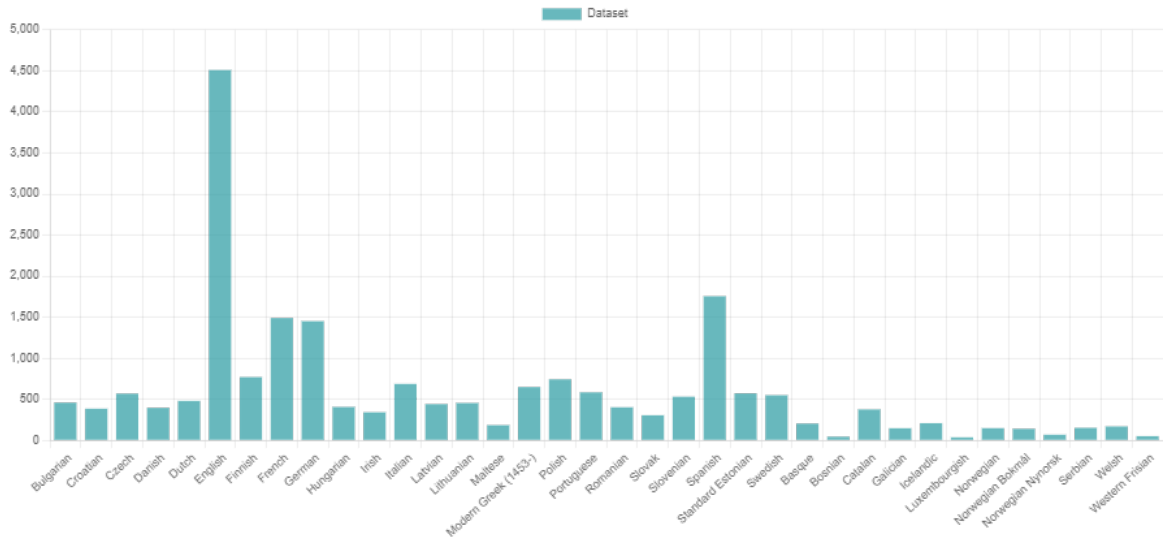

Figure 4: Number of multimodal datasets (i.e. media type: audio, video or image) available at the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones on 17.10.2022

big languages in various areas of Europe. Collecting resources and developing MT to translate between such languages is likely to result in large pay-offs for both sides. Finally, the potential of the language community as "data generators" is not to be neglected. Especially the speakers of smaller languages lend themselves towards leveraging of citizen science or crowd-sourcing approaches to data collection, dataset creation and tool evaluation. Crowdsourcing has proved to be an effective way of creating speech corpora, and many minoritised languages have benefited from a coherent pool of language activists who are keen to contribute to building new resources needed for LT and AI purposes.

However, in order for the above specific instruments for data collection and resource creation to be effective, the power of existing but unexploited data should be unleashed. There is much untapped, currently inaccessible data that could make a huge impact on the future of LT, if collected and applied appropriately. For example, there is a huge amount of aligned audio, video, subtitling and sign language data available in multiple languages already available in the archives of the public national broadcasters (radio, tv and national news agencies). High volume datasets are produced by the administration and other public institutions (e. g. in the domains of health, culture, media, justice or education), but they remain buried in untapped silos due to the reluctance of certain sectors of the Administration to effectively implement the European directives on open data and reuse of public information. Another type of data with untapped potential is the online user-generated content in the form of edutainment, influencer channels and vlogs. This data could be collected and processed to develop user-generated corpora necessary to build tools that could process modern written and spoken language.

Always ensuring a proper treatment of potential data bias in order to avoid pitfalls such as models making undesirable biased predictions that risk perpetuating gender roles, lead to unfair treatment of minority groups, etc.

An important prerequisite for the above is the adequate recognition of the importance and potential of language data. A future ELE programme should continue to raise awareness of the importance of language data. In particular, the following stakeholders, data holders and creators need to be targeted: the public sector, publishers, the language communities, and the national broadcasters.

4.2.2 Technology

Similarly to data, the identified gaps for technologies are extremely diverse across languages. While LTs for English are numerous and at the state-of-the-art, a number of very small minoritised languages, e. g. Karelian and Romani, lack very basic tools such as spell checkers. In the worst case they are not even recognised by operating systems. Nevertheless, there seems to be a generalised consensus that, when it comes to languages for which at least a minimum level of technological support has been achieved, the technologies most urgently needed are: discourse processing, bias detection and anonymisation, automatic subtitling, conversational systems and question-answering in the wider context of HCI, NLG (with summarisation mentioned quite frequently) and NLU. For instance, even English and German are currently supported by less then 100 HCI or NLG systems on ELG, while some languages like Bosnian and Norwegian Nynorsk are not supported at all (Figures 5 and 6).

Additionally, LTs should be developed and/or fine-tuned for new application areas and domains such as biomedicine, defense/security, media and government, smart homes and business processes support. It is moreoever considered critical to secure the presence of language-specific NLP modules in the major NLP platforms (commercial and non-commercial) such as spaCy, FreeLing, NLP Cube, TextRazor, Cloud Natural Language, Apache Open NLP, etc.

Recent advances in Cross-lingual Transfer Learning (CLTL), i.e. building of NLP models

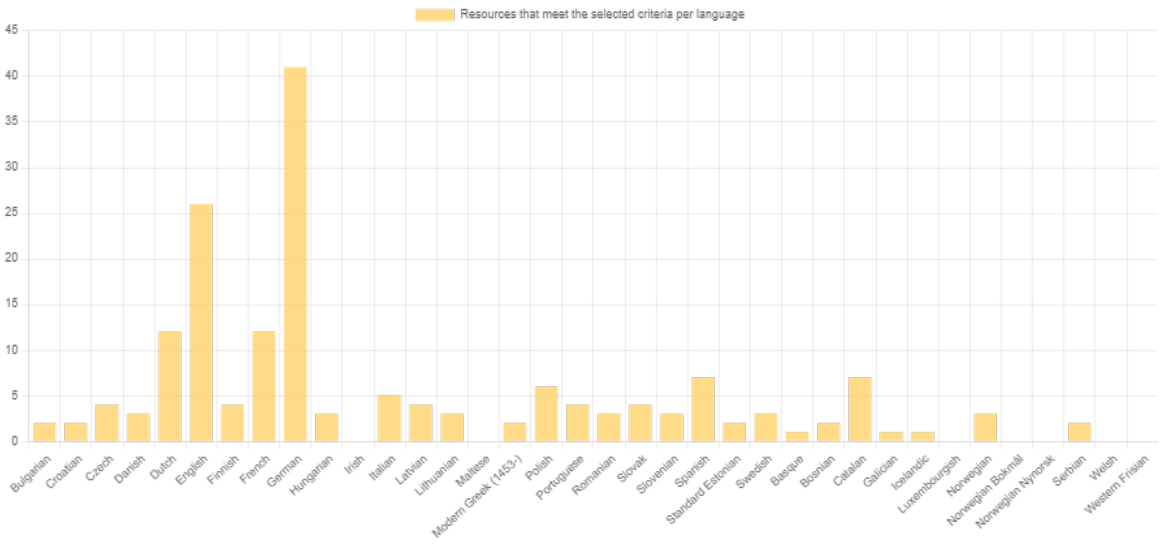

Figure 5: Number of Human Computer Interaction systems described at the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones on 17.10.2022

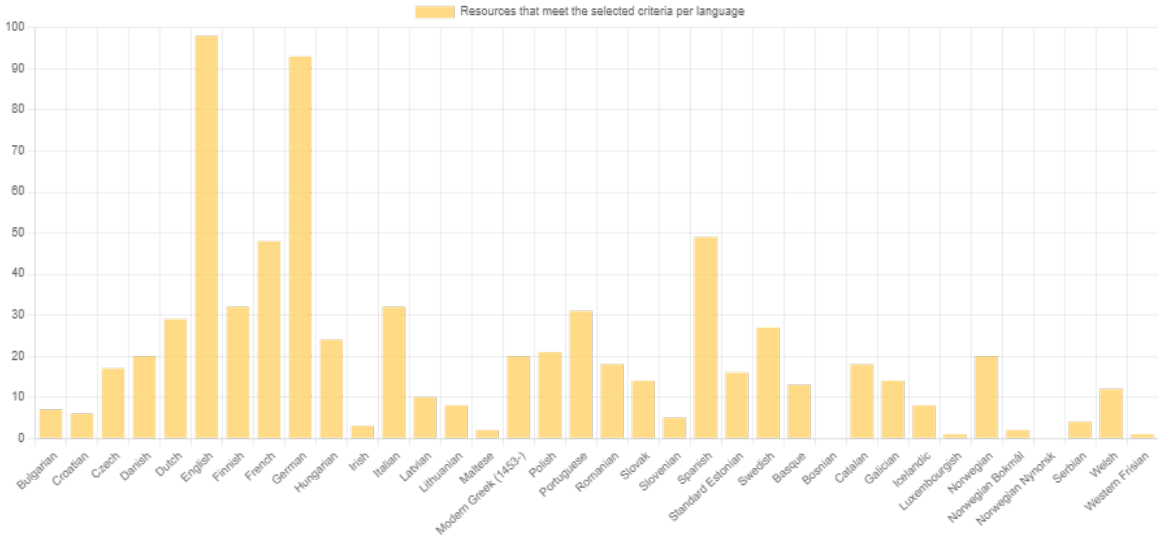

Figure 6: Number of Natural Language Generation systems described at the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones on 17.10.2022

for a low-resource target language by leveraging labelled data from other well-resourced languages such as English, have raised hopes for LT development for small languages. This points to a need for a general-purpose language-centric AI which can leverage cross-language and cross-domain resources and benefit from adaptation to local language varieties and specialised domains with small or medium-sized data sets. The application of such methodologies substantially cuts down the costs of developing cutting-edge LT for small languages.

The future of LT lies in connecting modalities (speech, text, image, video). – Multimodal fusion whereby different modalities of data – e. g. text, speech and image –- can entangle in complex ways so that rich multimodal processing of all sorts (including digital video, sound tracks, conversations, and virtual reality sessions) could form the bedrock of a new generation of content management technologies.

However, we should aspire to understand the internal workings of current language models better (in the spirit of the emerging research field “explainable AI”), in order to be able to exploit already existing linguistic knowledge (for instance, information about words collected in a lexical or conceptual resource) when training language models.

In addition, research should also address methods to potentially reduce their training data requirements, thus putting state of the art LT tools in reach of lower-resourced languages (including the official minority languages and widely-spoken immigrant languages).

There is, however, a critical caveat emphasised by the language experts: language-independent methodologies do not adequately respect the language-specific subtleties required for high-quality performance. Merely transferring technologies from e. g. English without adapting smoothly to another language and culture most often results in poor systems which are not fully functional and furthermore not inclusive to all parts of the society. Thus, despite the promise and prevalence of transfer learning, the need to invest in language- and culture-specific LTs is still present.

Finally, make sure that the necessary high-quality LT are widely available as a public commodity at scale.

4.2.3 Compute and Research Infrastructures

Access to advanced computing machinery and research infrastructures is a pressing demand. National and/or European coordinated actions should ensure flexible access to open high-performance computing research infrastructure. Especially as data continues to increase in scale and the need for training massive language models becomes more and more acute, it is essential to continue to support public computing infrastructures, with generous access rules and protocols for research organisations and for the European LT industry (e.g. startups and SMEs).

It has also been recognised that the creation of Language Resources Infrastructures (LRIs) that cater for storage, curation, and distribution of datasets and technologies/services, appropriately described with the relevant metadata and accompanied by clear and explicit licensing terms is a critical factor that benefits the overall availability of resources and tools, cultivates a data sharing culture, especially among non-LT communities, facilitates networking and knowledge exchange. Therefore it is highly recommended that a LT coordinated programme should ensure the maintenance, extension and sustainability of existing LRIs, such as the ELG and CLARIN. The existence of such LRIs alone, however, is not adequate. Clear policies and incentives for depositing at least all publicly funded LRTs in a LRI are necessary. Having a well-identified entry point for linguistic resources, models and tools, and their associated documentation, would be of great significance for many scientific disciplines, notably in the Social Sciences and Humanities (SSH) area, as well as for many industrial companies.

4.2.4 Situational context

Other policies and instruments that are recommended by the language informants are pertinent to a language's situational context, i.e. society, education, the legal framework, the role of the administration and the industry, etc.

Apart from English and a handful of other big languages, the biggest threat that globalisation of internet content and social networks poses for all other European languages is digital diglossia. Speakers of these languages, when going about their online lives, too often find it easier or even necessary to rely on other, more widely available, languages for determined services and information, because this gives them greater access to content and audience, and allows them to use more advanced technologies. This is true particularly for the younger generations, increasing the generational language gap and bringing the lesser-resourced language to digital extinction. It is also particularly true for bilingual societies, i. e. when two languages co-exist and/or they both share an official status. This prima facie case of linguistic inequality does not bode well for the outlook of Europe's cultural heritage. At the same time, the big corporations that currently dominate the LT market consider that there is no significant market demand for smaller languages, because their speakers are more or less served through other dominant languages. Consequently the industry does include small languages in their portfolios of innovative and popular AI applications, such as voice assistants, creating a vicious circle of lack of demand and lack of offer of LTs for most of Europe's languages. To tackle the above obstacles the context where a language thrives should encourage further use and it should facilitate deepening its penetration in digital life. More work must be done, e. g., to deepen a language's integration into social network applications, expand its use in business and employment-oriented services, and extend its reach into entertainment-related products. On the industry side, political pressure should be put on companies to support minority and small languages by opening up their platforms, e.g. by including LTs that have been developed by third parties. It could be in the form of a digital language technology act modelled on the anti-gatekeeping policies of the Digital Markets Act, and ensure that individuals or groups are not kept outside of the society at large, thereby avoiding the most serious threat to the future of minority and indigenous languages. From another perspective, investments on smaller languages could be required as part of a corporate social responsibility policy.

On a parallel line of argumentation it has been considered imperative that ownership of languages, including control of access and use, is transferred back to the language communities, and away from the major players. No single company alone can serve all languages, and no-one should expect them to, either. On the contrary, working solutions for most of the languages in the world will have to be developed by various third party groups: academics, open-source groups, voluntary organisations, language activists, SMEs, etc. The language communities should be mobilised and involved in all processes and projects on digitalisation and LT and they should embrace ownership of technologies for their languages.

A significant gap, concerning all areas of speech and language processing, is the scarcity of trained personnel and expertise. One of the reasons for this gap is that LT is an interdisciplinary field of research and, as such, it requires the combination of diverse knowledge and expertise. It additionally, and most critically, requires local expertise, i. e. knowledge of the target language. Therefore, traditional silos of learning (e. g. third level institutions, training programmes) need to adapt. It is imperative that the LT community expands, particularly involving human science departments at universities; this action requires promotion and lobbying on LT at political level. Even closer collaborations and communication between the "traditional" LT research community and the new AI field, e. g., through the establishment of dedicated academic LT training programmes is also recommended. Alongside the scarcity of LT experts, there is a risk of losing emerging talent to innovative power-players outside of Europe (with possibilities and salaries which can generally not be matched by Eu-

ropean players). Thus policies for retention of highly skilled LT professionals should ensure that talent is not only nurtured in Europe through education and training, but also retained.

A recurring issue pertinent to practically all European languages is the consideration of the legal status of LRTs. Issues related to IPR or GDPR render resource owners hesitant about sharing their datasets. In similar situations, non-explicit, unclear distribution and use terms restrict sharing, use and re-purposing of digital texts and language processing tools. The majority of resources – when made available – pose restrictions on the types of uses they allow (for example, for research purposes only or no derivatives), thus discouraging prospective users, hampering new research and development and leading to repetition in resources creation. Especially when it comes to applications that involve social media data, as they are often associated with delicate legal issues (related to proprietary rights or personal information), their dissemination and further exploitation potential is limited. This hinders the development of studies on opinion mining, fake news and hate speech detection, fact checking, on biases and ethical issues, to name a few. There are additionally major problems in getting the necessary authorisations to make use of language data in certain domains e. g. health and commerce, both of which are seen as sensitive. In practice IPR and GDPR often effectively block research access to language data.

It is therefore important that IPR and GDPR regulations become more flexible, allowing wider use of IPR protected data for the development of language technologies and resources and for research purposes in a way that does not harm the interests of the authors. Clear legal national frameworks and efficient transpositions and implementation of the European directives on open data are essential to ensure well-regulated access to language data for research and innovation (non-commercial and commercial) purposes.

Equally important is the provision of support and training to data owners on data and technologies licencing issues. What is more is the promotion of the adoption of standardised licences that are as open as possible, following the principle “Data should be as open as possible and as closed as necessary” and, in general, the promotion of an openness and sharing culture. Coordinated actions to promote the culture of data sharing should target all stakeholders, the public sector, research and industry, so that all potential data owners share with as few restrictions as possible. All public sector language data in particular need to be open and made available at least through the national open data portals. Openness can equally benefit all stakeholders, but most importantly it Support for open source solutions, which will allow small and medium-sized companies (and potentially also large ones) to develop applications without having to face the initial investment barrier.

An equally important issue highlighted by the ELE language informants, especially those representing small or minoritised languages, is the need to increase demand for and uptake of LTs for these languages. This can be achieved by measures for the enhancing digital literacy and in general up-skilling (minoritised) language communities, in an attempt to raise the level of the society’s ability to use the opportunities that language technologies have to offer. The administration’s and public sector’s role in this respect is considered critical. Although up to date for many languages the main force driving for LT development has been the public sector and state-funded projects have resulted in a great number of resources and tools, the public sector has almost exclusively acted as financier, but not as buyer or co-developers of LTs. To increase uptake, the administration should fully incorporate cutting-edge LTs, according to a programme linking AI and NLP with direct practical applications to eGovernment, thereby acting as a true driver of demand. Moreover, the administration should procure LTs that explicitly support the country’s/region’s language(s). Especially in bi-/multilingual regions, the administration should use its purchasing power to insist on provision of technology supporting in particular all languages used in the region.

One of the most significant shortcomings identified for the vast majority of languages is the lack of continuity in LT research and development support. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. In most countries,

there have been sporadic efforts, self-funded or partially supported within programmes in the wider IT or AI areas. This situation has resulted in discontinued and unsustainable resources and technologies, whereas it is unanimously agreed that targeted long-term support and funding streams, in the form of a coordinated ELE programme, are necessary in order to respond to the dynamic nature of digital technologies. Ideally, and in order to succeed in setting up such a programme targeted to LT, LT should be part of the national AI strategy and language policy: measures ensuring that the importance of language technology and language-centric AI is appropriately recognised should be included in the state policies for language, cultural and technological development.

Some of the characteristics of such a long-term ELE programme are the following:

- A coordinated ELE programme should be embraced and co-funded by national and regional governments, as appropriate and relevant, for instance when it comes to regional and minority languages or cross border collaboration for languages spoken in neighbouring countries.
- It should coordinate, align and synchronise regional, national and EU priorities and activities, especially with respect to research infrastructures and research priorities.
- With respect to the governance of the programme, responsibility should be assumed by a dedicated body at the national or regional level: A distributed organisation should overlook the implementation of the programme per country or language, while at the same time it holds responsibility for supporting and enrich education in the fields of LT and AI, increasing the visibility and ensuring the sustainability of existing and future LT resources, and facilitating improved knowledge transfer and collaboration between academic and industry stakeholders. Such an organisation can take the form for instance of a distributed centre of excellence.
- Develop standard formulations for public procurement to give the public sector the rights to language resources which emerge from translations and other services.
- It should equally fund research and the industry to tackle the identified disproportion between funding for research (TRL 1-4) and industrial activities (TRL 5-9).
- A shift in focus is required to recognise technology as an equally important axis for continued language use. This shift should see a broadening of scope in terms of funding within the wider lens of speech and language technology.
- Projects that are to be funded through an ELE programme need to be very carefully selected on the basis of their impact on the local economies and societies.
- An ELE programme should offer concrete opportunities for marketing and deploying LT applications.
- Knowledge transfer projects should be funded which will not aim at mirroring existing solutions for English but rather at supporting the development of adequate resources and tools for endangered languages.
- A legal framework to achieve digital inclusion should be created. Such a legal framework can be modelled on the Digital Market Act, and regulate access to language technology for all parts of a digital platform: from keyboards to digital assistants to localisation.
- in bilingual communities, prioritise support for the weaker language/variety (wrt. to regional/minority languages only): it must be a priority for decision makers to strengthen LT for the lesser used language to avoid weakening its equal status.

4.3 European Language Technology: The Voice of Europe's Citizens

In order to assess experiences and opinions of LT use among the end-users at large, an online survey was conducted that specifically targeted European citizens. The survey was designed with the purpose of taking into account the average citizen's opinions, individual needs, wishes and general demands, as well as to make sure that their voices play a decisive role in the pursuit of full DLE. This consultation with a larger and more diverse cohort of users and consumers allowed us to obtain an overall picture of the current scenario and future needs in terms of LT support across European languages. This also offers a representative basis for technological and scientific forecasting on how LTs can be deployed and applied in Europe by 2030 in the interest of DLE. To the best of our knowledge, this is the first ever survey focusing on LTs conducted on this scale, covering such a wide range of languages and such a large number of EU citizens.

4.3.1 Dissemination

The citzens' survey was hosted on the *QuestionPro*[71] platform, as it offered specific features required for the structure of the survey logic. The survey was launched in January 2022 and closed on May 1st, 2022. To ensure a wide reach across Europe, the first round of the dissemination process was carried out through Lucid's Survey Solutions[72] which offers online market tools and access to a large community of respondents world wide.

In order to make the survey widely accessible, it was made available in 35 languages and disseminated across 28 countries.[73] The survey was first set up in English and was automatically translated (where possible) using the *eTranslation* tool[74] and post-edited by native speakers of the target languages from the ELE consortium. After setting up the translated versions on *QuestionPro*, the same native speakers were requested to review a preview of the translated survey on the platform for the purposes of translation quality assurance. 28 of the translation target languages were those supported by Lucid. Following a request for volunteer translators from within the ELE consortium, six additional languages were included, where linguistic expertise was available: Bosnian, Icelandic, Luxembourgish, Macedonian, Maltese and Turkish.

For countries with more than one official language, we created a stand-alone version of the survey in each of the languages spoken in the country. For instance, in Spain, four different versions of the survey were set up in order to disseminate it in four languages, namely, Basque, Catalan, Galician and Spanish. Creating separate language versions of the survey for multilingual countries allowed us to specifically target regions in a country where communities of respondents that were speakers of that language were likely to be found. The Lucid responses were divided into quotas established by country, to ensure that the responses collectively provided a fair representation of European citizens, and the quota established for multilingual countries was divided between the languages spoken in the country. The sample size established per language was based proportionally on the size of the population speaking that language in the country.[75] In Spain, for instance, as Spanish is the most widely spoken language, 83% (750 responses out of 900) of the total quota was set up for the survey disseminated in Spanish, while the remaining 17% was distributed among the co-official languages of Basque, Catalan and Galician.

[71] http://www.questionpro.com
[72] https://luc.id, now known as Cint
[73] While ELE covers 85 European languages in total, we only produced translated versions for those languages for which native speaker post-editing was available. The 35 languages covered by the multilingual survey represent the support offered through the ELE consortium members.
[74] https://ec.europa.eu/digital-building-blocks/wikis/display/CEFDIGITAL/eTranslation
[75] Guidance on sample size to this effect was provided by Lucid, based on their previous similar campaigns.

For countries and languages not covered by Lucid's services, the survey was then disseminated via ELE partners and language informants. Through their professional and social media networks, they were able to target the communities of speakers of these languages. The regions and languages not covered by Lucid, but part of the ELE remit, included Luxembourg, Macedonia, Malta, Turkey, Iceland and Bosnia. The general survey link was also shared across the ELE partner network, from which a respondent could choose the version of the survey localised in their language. In total, 21,108 complete responses were collected through this online survey.

4.3.2 Analysis and Highlights of the Results

The European Citizen survey included a total of 11 questions, 6 multiple-choice questions, 4 single-choice questions and 1 open-ended question which allowed respondents to include any comments or feedback they had.[76] These 11 questions could be answered in approximately 5 minutes via computers or mobile devices. The full list of questions can be found in Appendix A.1. In order to ensure the reliability of the survey data captured, a number of data cleaning steps were required to remove responses that were deemed noisy or at risk of skewing the survey results. These steps are presented in more detail in Appendix A.2. The total number of valid responses that were used in the analysis of the data was 20,586 responses.

The demographic of the respondents is represented as follows:

27% of the respondents were between 25-34 years old. 23% accounted for both the 18-24 and 35-44 age brackets. The rest of the respondents were 45+ years old. 1% respondent preferred not to say. In terms of education, 35% of the respondents had reached High School level only, 23% held a Bachelor's Degree, 17% held a Master's Degree, with the rest reporting vocational training (11%), only some High School completion (7%) and holding a PhD (5%). 2% declined to say.

In the interest of space for this SRIA report, we limit our discussion of results here to three of the questions, which we believe to be of particular interest:

Question 1 *Please select all the words and terms you are familiar with or that you are able to understand right away*.

The respondents were presented with 10 frequently used terms in the LT space, along with an option to indicate that none of the terms were familiar. The purpose of this question was to gauge how much awareness the average EU citizen has of LT and its associated terminology. Figure 7 demonstrates that 'Machine Translation' and 'Chatbots' are most commonly understood terms amongst the respondents. More specific terminology describing the technology space ('Language-centric AI' and 'Natural Language Processing') or a newer technology ('Conversational Agent') is less commonly known. Interestingly, a non-negligible proportion of respondents (11%) are not familiar with any of the terms.

Question 6 *Please rate all the types of software applications, apps, tools or devices you use for your language(s). Tools you do not use for your language(s) do not need to be rated.*

The list of eight tools presented was: Search apps (e. g. Google, Bing); Personal assistant apps (e. g. Siri, Alexa); Proofreading apps (e. g. spelling and grammar checkers, autocorrect); Translation apps (e. g. Google Translate, DeepL); Automatic subtitling (e. g. news report, YouTube); Language learning apps (e. g. Babbel, Rosetta Stone); Chatbots (e. g. for customer support) and Screen readers.

[76] Note that this breakdown was incorrectly reported as 12 questions, 6 single-choice and 5 multiple-choice in D2.17.

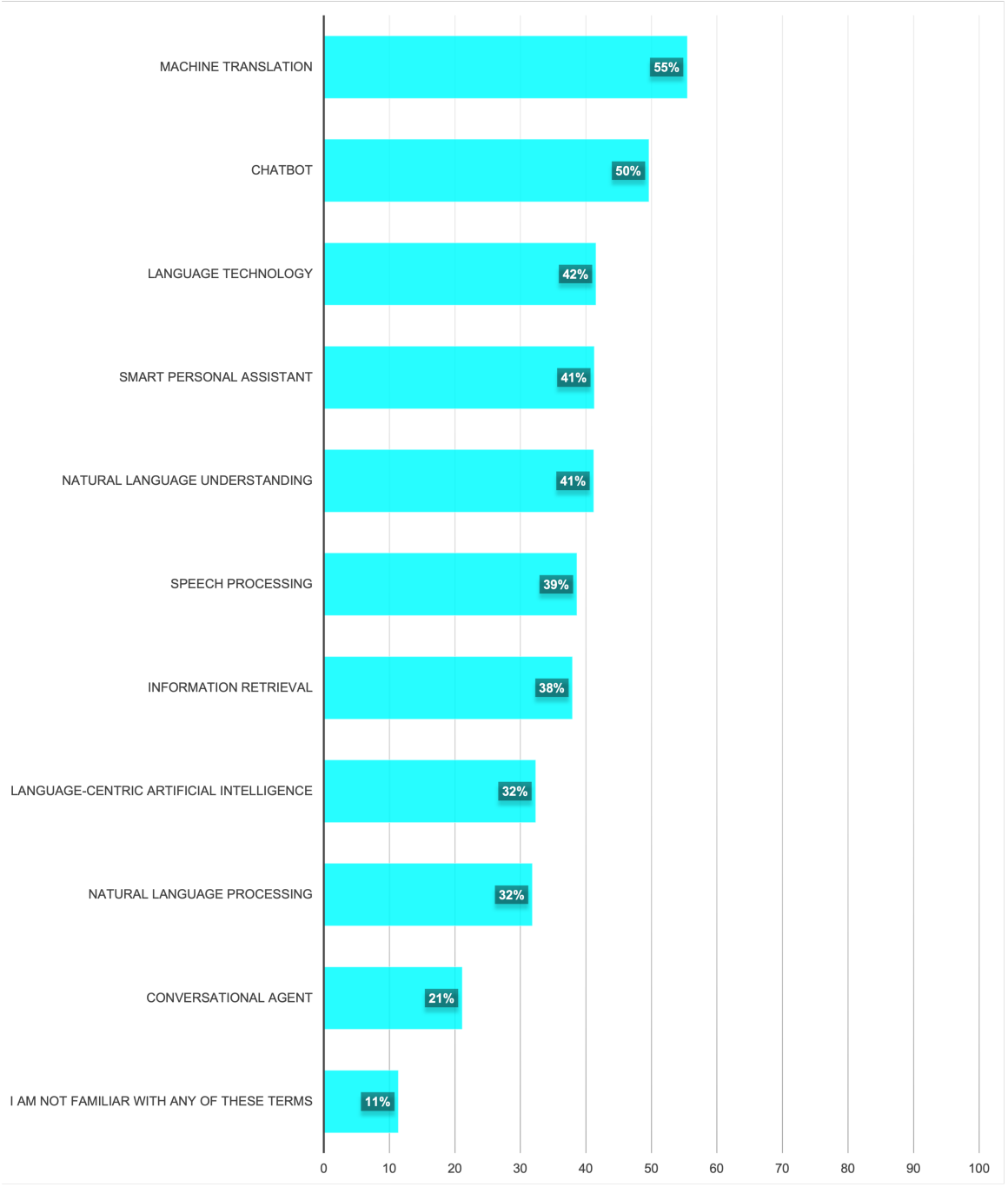

Figure 7: Responses to Question 1: *Please select all the words and terms you are familiar with or that you are able to understand right away.*

The ratings were based on a 5-point Likert scale. That is, the respondent had the option of rating 1-star (i.e. poor) through to 5-stars (i.e. excellent) for each of the eight tools presented, and for each language they selected in the previous Question 5. The aim of the question was to understand the perception of the average EU citizen and LT user of the quality of the tools

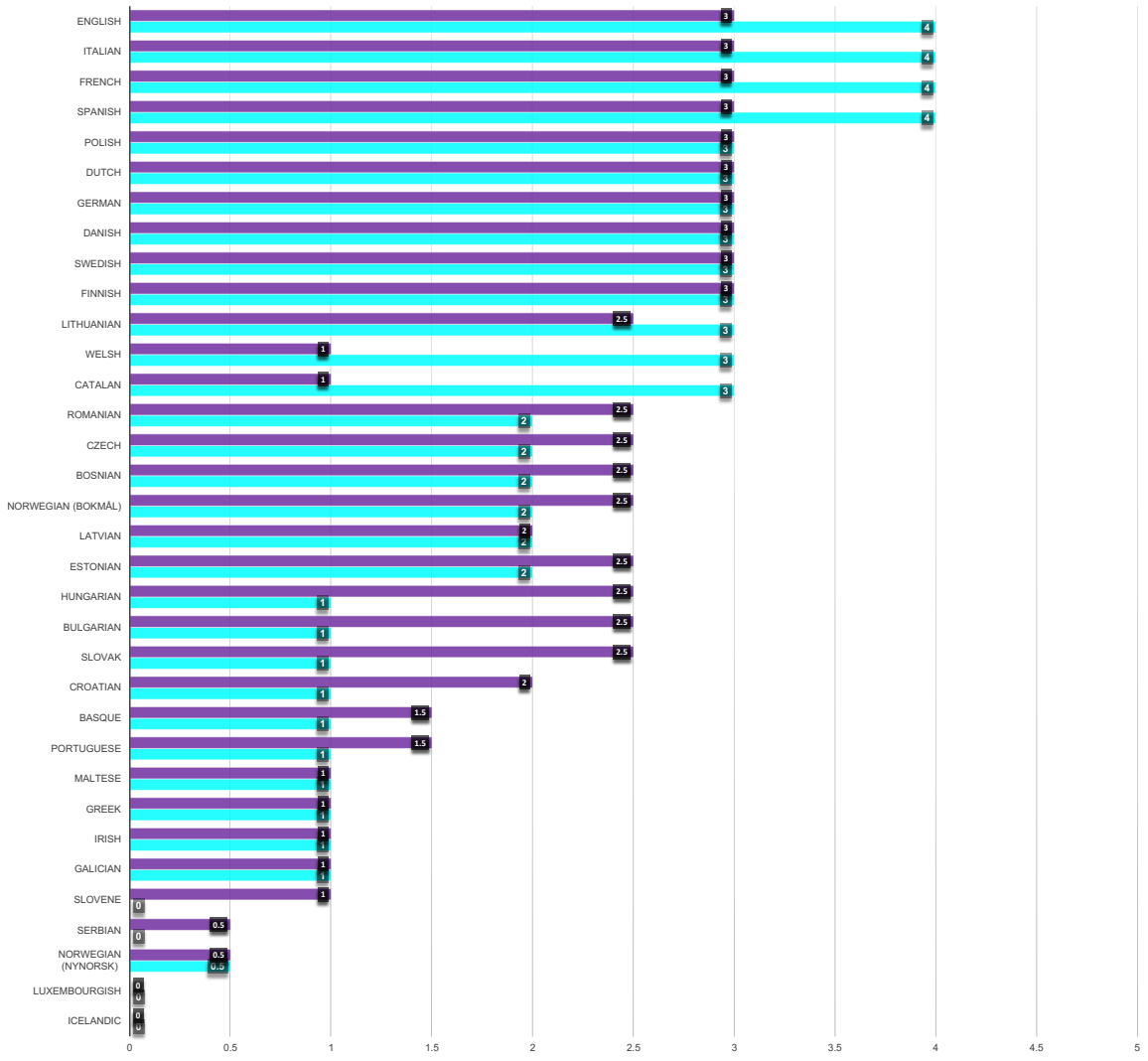

Figure 8: Responses to Question 6: *Please rate all the types of software applications, apps, tools or devices you use for your language(s). Tools you do not use for your language(s) do not need to be rated.*
Note that purple indicates the median score calculation and blue indicates the mode score.

that they use for each language they speak. It is important to note that the responses are subjective and limited to a respondent's awareness of how 'good' a technology has the potential to be. In the interest of space, Figure 8 presents only the languages for which language reports were produced in ELE's WP1 and only shows responses from the perspective of each language, as opposed to each tool. The calculations used for Question 6 results are explained in more detail in Appendix A.3.

To some degree, the results reflect the trend presented for the Technological DLE scores of the relevant languages, as shown in Figure 1, in terms of the quantification of the Technolog-

ical Factors of the DLE Metric.[77] The difference between the median score for English and the next well-resourced languages is not as stark, however. This could be explained by the fact that the ratings of the tools are bound to an upper limit of 5 and as a result the scores are 'flatter' and closer to each other. On the other hand, we can see that the mode score reveals that tools for English, French, Spanish and Italian received more frequent higher ratings. Nevertheless, the results provide a clear insight into the average European user's perception of LT quality.

Question 10 *What would be the top 3 advantages of improving apps and tools for all languages? Please select the three most important advantages in your opinion.*

The purpose of this question was to assess respondents' views on the benefits of LT. Notably, as seen from Figure 9, LT is regarded as key to enhancing multilingual societies from a linguistic diversity perspective. Of seemingly less importance to the average citizen is the economic advantage that arises from LT support.

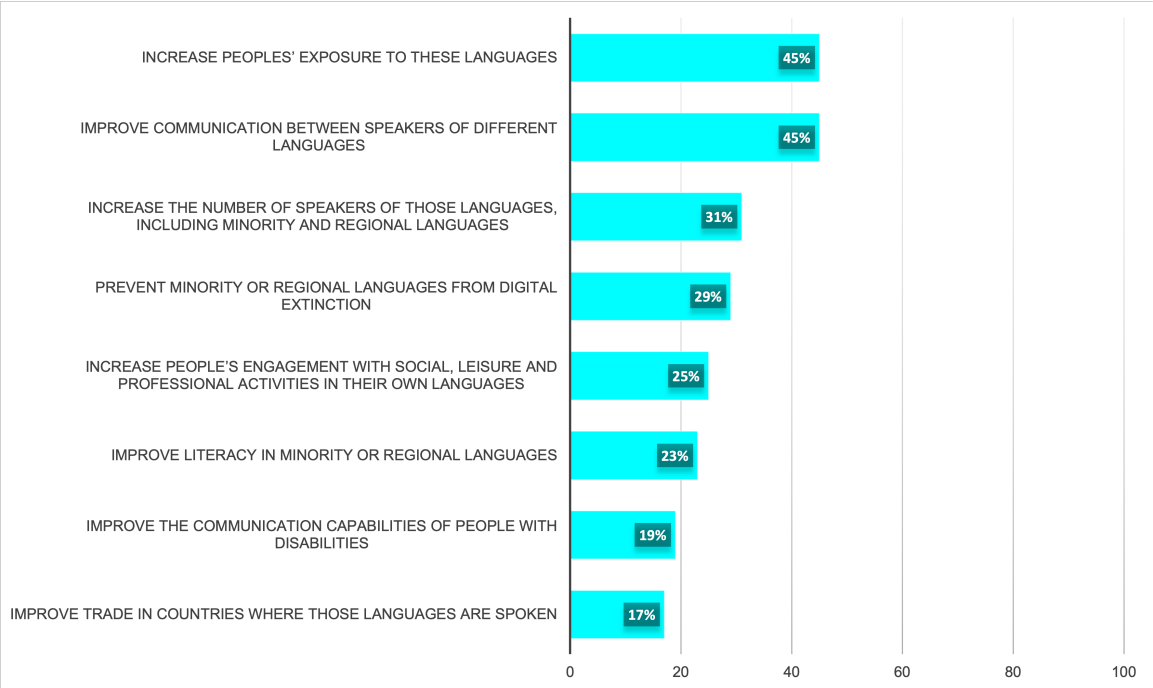

Figure 9: Responses to Question 10: *What would be the top 3 advantages of improving apps and tools for all languages?*

4.4 European Language Technology: National LT/AI strategies in Europe

All European countries consider AI an area of strategic importance. In December 2018, the EC and the Member States published the "Coordinated Plan on Artificial Intelligence", COM(2018)795, on the development of AI in the EU. The number of EU countries with an AI strategy (29 out of 30, 97%) demonstrates the success of the plan. Only Croatia has no official AI strategy as of yet. Figure 10 presents an overview of the LT funding situation in

[77] For the purposes of this specific analysis, which focuses on the availability and perceived quality of tools and applications, the situational Contextual Factors of the DLE Metric are largely irrelevant.

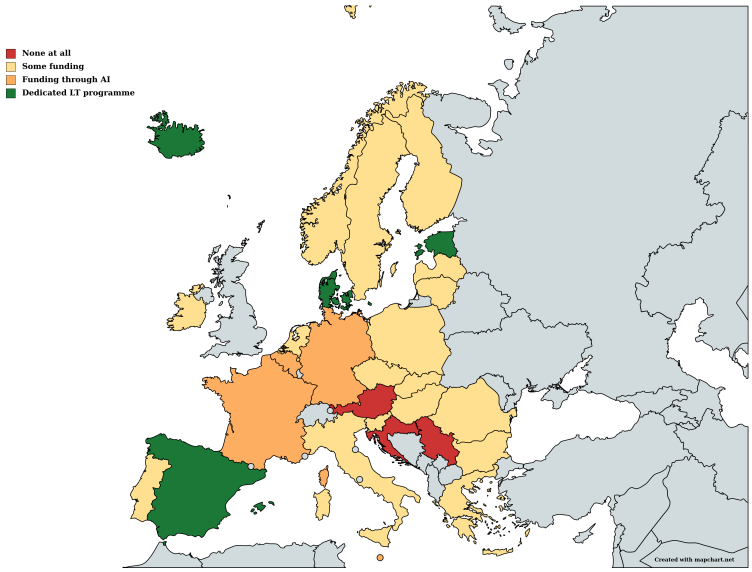

Figure 10: Overview of the LT funding situation in Europe

Europe. In green those countries with a dedicated LT programme. In orange those countries that explicitly provide funding for LT-related topics through AI. In yellow the ones with an AI strategy mentioning LT and finally, in red, those countries with no AI strategy or having an AI strategy but not mentioning LT at all.

The *AI Watch National Strategies on AI: A European Perspective in 2019* report also analyses the EU national AI strategies to identify areas for synergies and collaboration. It identifies several policy areas: human capital, from lab to market, networking, infrastructure, regulation. LT is mentioned as part of the Danish, Latvian, Maltese, Portuguese, Slovakian, Spanish and Swedish initiatives, so, as Rehm et al. (2020b) mention, LT is finally generating some momentum. LT is considered key, and language understanding is seen as one of the next generations of innovative AI technologies (STOA, 2017). For that, it is indispensable to set aside funding exclusively for LT. According to Rehm et al. (2020b), only four of the 30 surveyed countries do not have some type of LT funding. Four countries have programmes dedicated to LT (in green in Figure 10, Denmark, Estonia, Iceland, Spain), six provide funding for LT-related topics through AI (in orange in Figure 10, Belgium, Denmark, Estonia, France, Germany, Malta) and two (Ireland, Latvia) that do not have LT programmes, but rather a language strategy defined by their governments. The Spanish government has recently announced a new strategic project for economic recovery and transformation (PERTE in Spanish) called "The New Economics of Language". [78] The PERTE is presented as an opportunity to take advantage of the potential of Spanish and co-official languages as a factor for economic growth and international competitiveness in areas such as AI, translation, learning, cultural dissemination, audiovisual production, research and science. To do this, it has a budget of 1.1 billion euros of public investment, with the aim of mobilizing another billion in private investment. Moreover, following the lines of the Spanish Plan for the Advancement of LT[79], the Catalan government has also launched the AINA project,[80] the Galician government the Nós project[81] (de Dios-Flores et al., 2022) and the Basque Country the GAITU

[78] https://planderecuperacion.gob.es/como-acceder-a-los-fondos/pertes/perte-nueva-economia-de-la-lengua
[79] https://plantl.mineco.gob.es/Paginas/index.aspx
[80] https://politiquesdigitals.gencat.cat/ca/tic/aina-el-projecte-per-garantir-el-catala-en-lera-digital/
[81] https://www.xunta.gal/hemeroteca/-/nova/134792/xunta-usc-ponen-marcha-lsquo-proxecto-nosrsquo-que-permitira-incorporar-galego

project.[82] Currently, there are plans to create a network of research centers to coordinate the regional and national LT projects in Spain. Hopefully, these initiatives could also be used as a running model for the whole European ELE programme.

On 16 March 2022, the Committee of Experts of the European Charter for Regional or Minority Languages[83] adopted a statement on the promotion of regional or minority languages through artificial intelligence (AI).[84] In this statement, the Committee of Experts notes that AI applications may facilitate the daily use of regional or minority languages and support authorities in promoting them in accordance with the Charter. The Committee of Experts encourages states to promote the inclusion of regional or minority languages into research and study on AI with a view to supporting the development of relevant applications as well as to develop, in co-operation with the users of such languages and the private sector, a structured approach to the use of AI applications in the different fields covered by the Charter. The adoption of the statement was based on the study "Facilitating the Implementation of the European Charter for Regional or Minority Languages through Artificial Intelligence".[85]

In fact, the conclusions of the Education, Youth, Culture and Sport Council, 4-5 April 2022 call for the development of an ambitious digital policy for language technologies, translation and lifelong language learning and teaching. The EU wants to take advantage of new technologies to foster multilingualism, which nurtures cultural exchanges and facilitates access to culture.[86]

4.5 European Language Technology: SWOT Analysis

Taking into account all the reports, documents and national and international initiatives, this section summarizes the most relevant findings of these previous and existing reports analyzed here in terms of a SWOT analysis. It tries to identify the relevant internal and external factors that are favourable and unfavourable for creating an agenda and roadmap to make digital language equality a reality in Europe by 2030.

Table 3: SWOT Analysis

Strengths
- Emergence of powerful new deep learning techniques, tools that are revolutionizing LT. - Important basic LT has been developed, and applications that are used on a daily basis by hundreds of millions of users for speech recognition, speech synthesis, text analytics and machine translation are available. - Existence of multiple national and European LT research networks, associations, communities and other relevant stakeholders whose objective is to promote all kinds of activities related to research, development, education and industry in the field of LT, both nationally and internationally. - Existence of unique, valuable and potentially very useful data resources that can be exploited by current LT. An enormous amount of information is expressed in human language. - Increasing number of companies in LT and good level of readiness for the implementation of LT in production environments. - LT contributes to the development of inclusive digital societies, and is useful for digital transformation and responding to social challenges (accessibility, transparency, equity).

Continued on next page

[82] https://www.irekia.euskadi.eus/es/news/76846-gobierno-vasco-presentado-gaitu-plan-accion-las-tecnologias-lengua-2021-2024-cual-tiene-objetivo-integrar-euskera-las-tecnologias-linguisticas
[83] https://www.coe.int/en/web/european-charter-regional-or-minority-languages/committee-of-experts
[84] https://rm.coe.int/declaration-ai-en/1680a657ff
[85] https://rm.coe.int/min-lang-2022-4-ai-and-ecrml-en/1680a657c5
[86] https://www.consilium.europa.eu/en/meetings/eycs/2022/04/04-05/

Table 3 – *Continued from previous page*

- The European High Performance Computing Joint Undertaking (EuroHPC JU), a joint initiative between the EU, European countries and private partners, is developing a World Class Supercomputing Ecosystem in Europe that can alleviate the computing divide between large firms and non-elite universities. - A call for tenders has been recently launched for a 'Common European Language Data Space (LDS)'.
Weaknesses
- The LT markets are currently dominated by large non-EU actors, which do not address the specific needs of a multilingual Europe; Europe remains far behind compared to other developed technology-intensive regions of the world, on account of market fragmentation, insufficient funding and legal barriers, thus hindering online commerce and communication. Europe does not fully exploit its enormous potential in LT. - LT currently only plays a rather subordinate role in the political agenda and public debate of the EU and most of its Member States. Secondary topics are too dominant in the public discussion (for example, dangers of deep fakes). - There is a general misconception and over-hyping of the actual AI and LT capabilities, that can lead to exaggerated expectations and backfire. AI is often perceived in a polarized fashion as either "magical" technology that can solve any problem, or as a threat for jobs and workers to be replaced by machines. - While metrics and benchmarks exist for various sub-fields, it is often difficult for users or buyers to determine how well LT tools work with their own content. In terms of the nature of datasets used in benchmarking, businesses require realistic data and not "toy" examples that are not applicable to real-world problems. - No common EU policy has been proposed to address the problem of language barriers. GDPR/Copyright is a major barrier to the access and re-use of language resources, in competition with countries that adopt the "fair use" doctrine. - The Open Data Directive (2019/1024/EU) does not include language data as a high-value data category. Most of the data require extensive IPR clearing (to address Copyright and GDPR). - There is a lack of adequate LT policies and sustainability plans at the European and the different national levels to properly support European languages through LT. Only four of the 30 European countries studied have a dedicated LT national programme and only six have included LT funding through the AI national strategies. - Not all EU Member States are official full members of the CLARIN European Research Infrastructure. - There is scarce and limited LT support for non-official EU languages. No European LT association is represented in the new Data, AI and Robotics public-private partnership. - There is a lack of necessary resources (experts, HPC capabilities, etc.) compared to large US and Chinese IT corporations (Google, OpenAI, Facebook, Baidu, etc.) that lead the development of new LT systems. In particular, the "computing divide" between large firms and non-elite universities increases concerns around bias and fairness within AI technology, and presents an obstacle towards "democratizing" AI. - Compared to English, there are fewer LT resources and tools including language resources, annotated corpora, corpora covering various domains, pre-trained language models, benchmark datasets, software libraries, etc. There is an uneven distribution of resources (funding, open data, language resources, scientists, experts, computing facilities, IT companies, etc.) by country, region and language. - There is a weak open data sharing culture for many public stakeholders and SMEs across a number of EU member states, due to the lack of awareness or implementation of the EU Open Data Directive. - The investment in AI does not reflect the real importance of LT.

Continued on next page

Table 3 – *Continued from previous page*

There is a fragmented European market with an extremely large and varied base of more than 1000 SME companies that develop LT. Small to medium national technology companies have little capital and investment in LT capabilities. The markets are small for low-resource language speakers. - In many countries, there are weak links between academia and industry and insufficient effective mechanisms for knowledge transfer. There is weak internationalization of R&D&I and innovation.
Opportunities
- Many new powerful monolingual, multilingual and cross-lingual deep learning LT capabilities are available. - LT is key for the realisation and support of European multilingualism. - LT is used in practically all everyday digital products and services, since most use language to some extent, especially all internet-related products such as search engines, social networks and e-commerce services. - LT can impact on sectors of fundamental importance to the well-being of all European citizens, such as health, administration, justice, education, culture, tourism, etc. LT offers effective solutions to facilitate monolingual and multilingual communication, also for the deaf and hard of hearing, the blind and visually impaired and those with language-related disabilities or impairments. - LT is one of the most important AI application areas with a fast growing economic impact. Enormous growth is expected in the global LT market based on the explosion of applications observed in recent years and the expected exponential growth in unstructured digital data. - Europe can play an economic leading role with its neighboring countries through good partnerships based on the use of LT customized to other languages. - Growing trend for the LT market and industry in Europe regarding the exploitation of digital resources and data of linguistic interest. Digitisation is one of the key means to generate new economic growth. - Consolidation of a competitive LT industry that harnesses the potential of research and academia both in educating well-trained LT professionals and in transferring research results to industry and public administrations. - Increasing interest in higher education to organise Bachelor and Master in Science degrees (BSc, MSc) level education in AI/LT. When coordinated and quality-checked carefully, this could lead to an important increase of the AI/LT-educated workforce. - Increasing awareness about the possibilities of AI and LT and the necessity to invest and coordinate efforts. - Increasing awareness of the need to cultivate a culture of data sharing in public administrations will provide the foundations upon which public sector LT can thrive. - Substantial breakthroughs and fast development of LT offer new opportunities for digital communication; current multilingual and cross-lingual deep learning LT allows for the creation of new multilingual pre-trained language models and systems that can leverage and balance LT across all European languages. - Community-driven open sourcing of large models happens at breakneck speed, empowering collectives to compete with large labs. - Openness of infrastructures for data and technologies is improving.
Threats
- As shown previously in this report, at least 18 European languages are still in danger of digital extinction, thwarting the fundamental concept of the languages of Europe being equal. - A divide will inevitably emerge across EU Member States, as countries with sufficient language technologies will gain economic advantage.

Continued on next page

Table 3 – *Continued from previous page*

- The lack of digital support for a language can, over time, lead to (1) users defaulting to writing or speaking in another supported language or (2) users ceasing to interact with technologies. In the case of (1), this is a clear step towards language shift and eventual language decline, particularly amongst younger generations. In the case of (2), a divide arises in levels of information accessibility and economic progression across language communities. - Deep learning LT and large pre-trained language models have shortcomings and limitations. Large language models have limited real-world knowledge, can generate biased and factually incorrect text, may contain personal information, etc. They are also expensive to train, difficult to interpret or explain and have a very heavy carbon footprint. It is important to understand the limitations of large pre-trained language models and put their success in context. - AI is a very broad area, which overshadows and dwarfs the importance, benefits and contributions of LT, especially in Europe. - Loss of LT skills and human capital trained in Europe due to the lack of sufficient research, transfer and funding opportunities. - Inability to attract or retain EU researchers and experts skilled in LT and AI. - Most work in the LT ecosystem requires expert-level skills in the realm of tools related to data management, data science and NLP processing. This creates bottlenecks in industry since it does not allow domain experts (e. g. experts in finance) to become actively involved without rather extensive tool training. - Growing development of the sector in US and China that will sooner or later penetrate the European application market, limiting the Digital Language Equality opportunities as described in this report. - The complexity of copyright/GDPR/Open Data directives makes the access to resources too costly, unclear and risky. - Fear of many jobs becoming redundant due to the deployment of AI-powered technologies. Without clear LT strategies, policies and standards, interoperability issues will arise across the various data formats and system intfrastructures developed at an enterprise level. - Multilingual countries often feature a more dominant language that influences the language medium through which education is offered across society. While lesser-spoken language-medium schools are key to ensuring continued use of the language across generations, the availability or lack of language technology to support learning could eventually create a divide in the levels of education on offer to citizens, contributing further to societal inequalities. - Online data mining is often used by governments and media to gather information on events, political issues and public sentiment. However, in a multilingual society, only the opinions or comments of those in the technologically supported languages will be represented. In other words, the voices of many will be left unheard, unrepresented and unaccounted for. - In multilingual environments, it is common to find online multilingual text known as code-switching. Most LT tools are not equipped to deal with this naturally occurring linguistic phenomenon and as a results, such text is ignored or overlooked. - Scaling laws for large language models refocus on amount and variety of data which could be a problem for less-resourced languages.

5 Digital Language Equality in 2030: The ELE Technology Vision and Priority Research Themes

The ELE Programme, specified in the form of this SRIIA, will serve as the blueprint for achieving DLE in Europe. While the political and societal goal is indeed reaching *full Digital Language Equality across all European languages* (and, at the same, preventing digital extinction of many of our languages in Europe), the scientific goal envisioned to be reached by 2030 is *Deep Natural Language Understanding*.

Human languages are incredibly complex. We do not yet have algorithms or machines that are able to accurately and seamlessly integrate modalities, situational and linguistic context, general world knowledge, reasoning, emotion, irony, sarcasm, humour, metaphors, culture or explainability, or that are able to do all of this as required on the fly and at scale reliably across domains for the many languages of Europe and beyond. All of these bear on and are the hallmarks of truly *deep* language understanding in contrast to shallow processing (in the sense that the resulting application using LT is not an opaque black box but able to explain itself), meaning: *Why* did the system make the decision it made given the linguistic, situational or communicative context (across modalities), linguistic knowledge or general world knowledge?

Since 2010, the topic has been receiving more and more attention, recently also increasingly on a political level. In 2017, the study *Language Equality in the Digital Age – Towards a Human Language Project* (STOA, 2017), commissioned by the European Parliament's Science and Technology Options Assessment Committee (STOA), concluded that the topics of LT and multilingualism are not adequately considered in current EU policies. Over the coming years, AI is expected to transform not only every industry but society as a whole. The scientific and technological roots of LT are deeply embedded in AI and Computational Linguistics, especially with regard to the development of knowledge-based systems for language understanding. An increasing number of researchers perceive full language understanding to be the next barrier and one of the ultimate goals of the field at large and the next generation of innovative AI technologies. The European Parliament adopted, on 11 September 2018, with a landslide majority of 592 votes in favour, a resolution on "language equality in the digital age" (European Parliament, 2018) that also includes the suggestion to intensify research and funding towards Deep Natural Language Understanding. Both the STOA Report and the EP Resolution emphasise the enormous need for a large-scale, multidisciplinary LT development and deployment programme that benefits European society, industry and politics. The technological, societal and economic opportunities of developing technologies for cross-lingual and cross-cultural communication in Europe, and beyond, are almost endless.

The internal and external consultations and surveys conducted by the ELE consortium as well as the empirical results as exposed by the ELE Dashboard confirm that there is still a huge gap in LT support for English and all other European languages, with dramatic differences in several cases. Even though there is an increased interest in closing the gap and in developing more technological support for under-resourced languages, limited funding, uneven demand and various obstacles make it a very challenging endeavour. Basic research is still urgently needed. For many languages there is a severe lack of available data. The fragmentation of the LT industry remains a serious hindrance. On the other hand, the last decade has seen progress on a larger scale than could have been imagined 10 years ago. Many experts highlight European excellence, also on a global level and consider leadership in LT and language-centric AI to be possible if the necessary conditions are created by political decision-makers.

While the goal of Deep Natural Language Understanding by 2030 is ambitious, it can be reached by setting up a shared programme between the EU, the Member States, regional, national and international authorities and other stakeholders, including industry. It must necessarily include a balanced mix of basic research, applied research, technology development, resource development, innovation and commercialisation, driven by regional and national funding as well as funding from the European Union; education and talent retention must be taken into account, too, to ensure long-term sustainability. The programme should run for approx. ten years, so that the political and societal goal as well as the scientific goal can be adequately addressed. Public procurement and a policy change towards "LT enabled multilingualism" are crucial related aspects.

5.1 Priority Research Themes

Natural language is at the heart of human intelligence. Languages are the most common and versatile way for humans to convey and access information. We use language, our natural means of communication, to encode, store, transmit, share and manipulate information. As mentioned in chapter 3, unstructured data independent of it's modality is the usual case (around 80%) when dealing with digital information in multiple languages.[87] That generates a huge challenge for any organisation that wants to exploit and process its information given that most compter systems today process dominantly structured data. The lack of structure causes ambiguity and complexity in the processing and requires knowledge about the context and the world. Therefore, language is and must be at the heart of our efforts to develop AI technologies - as well as simply many traditional IT systems.[88] No effective AI-powered tool can exist without mastery of language.[89] Thus, language is the next great frontier in AI.[90] In fact, currently, LT is arguably the hottest field of AI.[91]

Despite claims of human parity in many of the LT tasks, Natural Language Understanding (NLU) is still an *open research problem* far from being solved since all current approaches have *severe* limitations. The development of new LT systems would not be possible without sufficient resources (experts, data, computing facilities, etc.). Creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application is needed, to foster technological progress, while encouraging deeper understanding of the mechanisms by which they are achieved. All these efforts will then lead to long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication, to create transparent digital language equality in Europe in all aspects of society, from government to businesses to the citizens.

5.1.1 Overall Goal: Deep Natural Language Understanding

Much has been said in recent times about the expected impact of intelligent systems in many aspects of our lives. Today's large amount of available data, produced at an increasing pace and in heterogeneous formats and modalities, has stimulated the development of means that extend human cognitive and decision-making capabilities, alleviating such burden and assisting our drivers, doctors, teachers and scientists, and sometimes even replacing them.

In scientific disciplines like biomedical sciences, some like (Kitano, 2016) even propose a new grand challenge for this kind of systems: to develop an artificial intelligence that can make major scientific discoveries and that is eventually worthy of a Nobel Prize. Though still far from realization, this scenario suggests the time is ripe for a shared partnership with machines, whereby humans can benefit from augmented reasoning and information management capabilities if machines are endowed with the necessary intelligence to assist with such tasks. Through such partnership, we can expect a virtuous circle of training data collection, active learning, and interactive feedback, which will result in self-adaptive, "everlearning" systems. We have already seen signs of such partnership, for example in the application of generative language models like GPT-3 to produce text given a prompt, with applications in a wide variety of business sectors. Based on these developments, some suggest[92] that the future of artificial intelligence lies in the development of systems that allow maintaining a conversation with a computer. This scenario goes beyond current chatbot technologies, which many deemed as mere digital parrots, able to copy form without understanding meaning,

[87] http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/
[88] https://hbr.org/2022/04/the-power-of-natural-language-processing
[89] https://www.nytimes.com/2022/04/15/magazine/ai-language.html
[90] https://www.forbes.com/sites/robtoews/2022/02/13/language-is-the-next-great-frontier-in-ai/?sh=6995a0865c50
[91] https://analyticsindiamag.com/is-nlp-innovating-faster-than-other-domains-of-ai
[92] https://www.theverge.com/22734662/ai-language-artificial-intelligence-future-models-gpt-3-limitations-bias

but nevertheless capable of creating a dialogue with the user. This is something that often seems missing from the introduction of AI systems like facial recognition algorithms, which are imposed upon us, or self-driving cars, where the public becomes the test subject in a potentially dangerous experiment. With AI writing tools, there will be the possibility for a conversation. However, this will require advances in knowledge representation, true understanding of meaning and pragmatics, and the ability for the models to reason, deduct new knowledge and relations, as well as explain and interpret their predictions in ways that humans can understand and relate to. The artificial intelligence community and particularly the areas related to text understanding will soon need to address issues other issues like fairness in ways that tangibly and directly benefit disadvantaged populations. We have spent large amounts of effort discussing about fairness and transparency in our algorithms. At the algorithmic level, fairness has to do with the absence of bias in the models that e.g. in natural language understanding are used to address tasks that may range from the evaluation of mortgage applications or insurance policies to medical examination and career recommendation. If the algorithms are biased, then so will the outcome of their predictions be and inequalities would be perpetuated as the use of artificial intelligence unrolls more and more deeply in society. This is essential work, but now it is time to develop systems and tools that have a tangible impact in business and society. The lack of resources in a specific language to train a natural language understanding model in such language is another source of discrimination. A very visual example in a related domain has to do with the use a smartphone navigation app in a wheelchair — only to encounter a stairway along the route. Even the best navigation app poses major challenges and risks if users cannot customize the route to avoid insurmountable obstacles. Similarly, the lack of availability of service functionalities in all language will have an undesired effect in the respective populations. Accessibility, education, homelessness, human trafficking, misinformation, and health among others are all areas where artificial intelligence and text understanding can have a major positive impact on people's quality of life. So far, we have only started to scratch the surface.

All of the above - specifically in Europe - has to be put in the multilingual context to assure digital language equality among all relevant languages. In addition, where appropriate, translation (both of spoken and written language) is still highly relevant, in terms of quality (including specialized domains), speed, footprint and accessibility.

Therefore, the specific areas to concentrate on during the next ten years are

- Machine Translation, for both written and spoken language;
- Text analytics from basic language processing (BLARK) to information extraction and grounding; natural, unbiased and context- and culture-aware text generation;
- Speech processing in all aspects, from language identification to high-quality ASR in adverse environments and natural TTS;
- Data and Knowledge acquisition, curation, persistence and standardization, across all languages;
- Infrastructure development as a basis for all data, tools and services.

These areas are described in more details below, including some specific recommendations related to their feasibility, achieving the progress needed, and way of supporting them.

5.1.2 Machine Translation

Multimodal MT A new definition of how context-aware MT should be addressed is essential, including understanding the context-related issues in different languages and domains,

and the context span necessary to solve those issues, including external information (meta data). This external information can go beyond text data and include images, videos, tables, etc. by developing **multimodal MT systems** (Yao and Wan, 2020). Future systems should combine different sources of information to help disambiguate words in the descriptions or reviews of products for e-commerce or online shopping, etc. In this line of research, for the 70 million Deaf people on the planet, improved **sign-language translation** is needed to enable them to communicate with one another as well as with non-Deaf communities. As well as many other things in this space, formalisms that capture manual and non-manual features and combine them together would be a significant contribution. Similarly, there is an expanding preference (especially among younger users) for voice-based interaction with devices, which points to more and more applications for **speech-to-text** and **speech-to-speech translation**. By 2030, it is likely that the automatic speech recognition-MT-speech synthesis pipeline will have been replaced by more direct approaches that model spoken language translation as an end-to-end process (Di Gangi et al., 2019). However, clearly more work needs to be done in this regard.

European MT Sovereignty Future publicly available MT systems should not rely or depend on large non-European companies. Otherwise there is a serious risk that what is freely available now (e. g., Google Translate, Bing Translator, etc.) could (easily) be taken away if those companies find a way to increase revenue in other directions, so that they deprecate their MT offerings, as has frequently happened with other services and technologies developed and provided by these large corporations.

MT evaluation The MT community still largely relies to a large extent on one of the first automatic MT quality evaluation metrics, BLEU, and there is a noticeable reluctance to abandon this measure despite a large body of research pointing out its drawbacks (Mathur et al., 2020; Kocmi et al., 2021). Future MT systems should be evaluated by new automatic metrics which represent better approximations of human judgments and also ideally abandon the dependence on human reference translations, which is a serious limitation. While the quality of human translations can be revised and controlled in the future as mentioned above, the problem of scores providing information only on how close the system output is to just a single possible correct translation among many does not essentially reflect any actual aspect of translation quality in the real world. Future metrics should be designed in a flexible manner so as to use the original text and MT output to provide information about the desired quality aspects for the specific task at hand. Another related aspect to be prioritised concerns the need to extend automatic MT quality evaluation from the currently dominating sentence/segment level to the entire text that is processed with machine translation, to adequately capture suprasentential discourse phenomena that are typically disregarded at present (e.g. anaphora, textual cohesion and coherence, cross-references, etc.).

5.1.3 Text analytics and TDM

NLP with common sense and reasoning Integrating common sense and reasoning in NLP systems has long been seen as a near impossible goal – until recently. Now, research interest has sharply increased with the emergence of new benchmarks and language models (Mostafazadeh et al., 2016; Talmor et al., 2019; Sakaguchi et al., 2020; Ma et al., 2021; Lourie et al., 2021). This renewed interest in common sense is encouraged by both the great empirical strengths and limitations of large-scale pretrained neural language models. On one hand, pretrained models have led to remarkable progress across the board, often surpassing human performance on leaderboards. On the other hand, pre-trained language models

continue to make surprisingly basic and sometimes bizarre mistakes.[93] This motivates new, relatively under-explored research avenues in common sense knowledge and reasoning.

Alternatives to data-intensive NLP The data-hungry and blackbox nature of current deep-learning approaches is leading to potential negative impacts on the environment (carbon footprint of processing power) and challenges in interpretability. There is therefore a clear need for further research into less computationally-intensive and less opaque solutions for TA and NLU. One growing avenue of research is the combination of language models with symbolic approaches (knowledge bases, knowledge graphs), which are often used in large enterprises because they can be easily edited by human experts, even though this is a non-trivial challenge. As such, there is much value in investigating possible opportunities to leverage both structured and unstructured information sources, and to enhance contextual representations with structured, human-curated knowledge (Peters et al., 2019; Colon-Hernandez et al., 2021; Lu et al., 2021). Other promising areas of research to overcome the need for large datasets include few-shot or even in zero-shot approaches including natural language prompting (Brown et al., 2020; Ding et al., 2021). *Prompting* or prompt-based learning is a technique that involves adding a piece of text (called *promts*) to the input examples to encourage a language model to bring to the surface the implicit knowledge that is required, thus helping the language model to perform the task at hand. The application of zero-shot to few-shot transfer learning with multilingual pre-trained language models, prompt learning and self-supervised systems opens up the way to leverage NLP techniques for less technologically supported languages, to promote DLE in Europe by 2030.

Human-in-the-loop NLP Traditional linear NLP development pipelines are not designed to take advantage of human feedback. Advancing on the conventional workflow, there is a growing research body of *Human-in-the-loop (HITL) NLP* frameworks, or sometimes called *mixed-initiative NLP*, where model developers continuously integrate human feedback into different steps of the model deployment workflow. This continuous feedback loop cultivates a human-AI partnership that not only enhances model accuracy and robustness, but also builds users' trust in NLP systems (Wang et al., 2021). This form of human intervention when developing NLP tasks is not new, and has been used for a long time e. g. in MT, where automatically generated translations are often post-edited by humans, who improve upon the raw output of the system. In areas such as text simplification or summarisation, machines have shown a reasonable capacity to recognize the most salient points of large documents, but have problems turning these points into coherent texts. But having a machine highlight the key ideas in a document and a human turn that into a short snippet, outperforms either approach working alone. Further research into this area is required as AI and NLP become embedded in everyday work processes, leading to increased human-computer interaction.

5.1.4 Speech

Broader voice coverage and diversity The issue of the lack of availability of data affects all 4 research themes. However, in the case of Speech Technology, the amount and breadth of coverage of training data has a direct impact on accessibility and the potential user base of a given speech-based technology. This not only includes language coverage (which limits current STs to speakers of certain well-supported languages) but also coverage of speakers of dialects, non-native speakers with pronounced accents, voices of the elderly, children's voices, or voices of the orally impaired. For example, the diversity of contexts and speakers

[93] https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

represented by popular ASR benchmarks like Librispeech (Panayotov et al., 2015; Garnerin et al., 2021) (read speech), and Switchboard (Godfrey et al., 1992) (spontaneous speech) is limited. Recent works attempt to address this problem by introducing benchmarks that mimic real-world settings, with the goal of detecting model biases and flaws (Riviere et al., 2021). The results obtained on this set show that while contemporary models do not appear to have a gender bias, they often reveal significant performance differences by accent, and much greater differences depending on the socio-economic background of the speakers (Backfried et al., 2022). Speech datasets therefore need to be carefully collected and curated to avoid marginalisation of subgroups of the population who otherwise would be denied access to advances in this field. Where such data is limited in availability, new methodologies for augmenting these datasets also need to be explored.

Multi-speaker environments While most research to date has focused on a single user's interactions, speech technologies embodied in virtual assistants are becoming increasingly popular in social spaces. This highlights a gap in our understanding of the opportunities and constraints unique to multiple user scenarios. These include detecting if users address the system or other participants, speaker diarization (see Park et al. (2022) for a review of recent advances in speaker diarization with deep learning methods), understanding aspects of social dynamics, and finding interaction barriers are some of the factors that restrict the usefulness of voice interfaces in group settings (Backfried et al., 2022). Similarly, Kanda et al. (2021a) and Kanda et al. (2021b) have demonstrated examples of research on E2E multi-speaker ASR for meeting transcription. Multi-speaker, multi-channel, multi-microphone setups may provide further angles of research and lead to improvements in this area.

Evaluation metrics and accuracy The single most frequently mentioned hindering factor for the broad adoption of speech technology is one that has been discussed for the past 40 years: accuracy. The perceived accuracy and its exact meaning have changed dramatically – from individual words being misrecognised, to intentions that are not correctly interpreted in complex situations, with the need for accuracy assessments reaching well beyond the actual accuracy of an ASR system only, regarding it in a more comprehensive and embedded manner. Whereas WER as an evaluation measure has had its merits to measure progress in ASR (and still does so), more comprehensive approaches to measuring the impact of ASR performance on downstream tasks and actual deployments may require novel approaches. WER alone clearly does not provide the full picture when it comes to the perceived performance and usability of complete systems comprising several kinds of speech and language technologies (Backfried et al., 2022). Similarly, the most popular measure of quality in TTS is the Mean Opinion Score (MOS) (Goldstein, 1995; Rec, 2006). Considerable time and effort must be devoted to the development of this subjective evaluation, as a large number of individuals is needed to reliably rate the TTS systems. Although MOS tests are still the most frequently used option to assess TTS, they have been criticised as they offer only a general measure of the overall quality and may not be suitable for evaluating long synthetic speech passages (Clark et al., 2019; Wagner et al., 2019; Backfried et al., 2022).

5.1.5 Language Data, Resources and Knowledge

Data and Knowledge Resources Availability The availability of suitable language data for use in both training and evaluating today's state-of-the-art data-driven LT tools is crucial. In particular, when it comes to current deep-learning paradigms, the size of a training dataset directly correlates with the quality of the LT tool. However, the current lack of parity in such resources for different languages translates directly to digital language inequalities. Across the EU, language data availability varies, often due to the number of speakers, which can

determine the amount of commercial interest and the extent of digital content creation. In addition, the willingness of enterprises or public sectors to make data available plays a role along with restrictions through GDPR, licensing or copyright. Furthermore, there is much untapped potential in terms of quality language data across EU public sectors. In fact, in relation to MT data, Berzins et al. (2019b) report on the difficulties experienced across a number of EU member states in accessing public sector language data – due to the lack of awareness or implementation of the EU Open Data Directive. In general, the value of language data is still widely unknown. To add to this data scarcity, many LT tools require knowledge databases or annotated or labelled data, which can be a time-intensive task that often requires skilled domain expertise, and which is a costly overhead for both the research and industry communities. Moreover, the lack of coverage in particular domains poses restrictions for advancements in LT in these areas: e.g. medical, health, pharmaceutical, legal, finance, insurance, science, manufacturing, publishing and so on. Research is thus needed to find faster, cheaper, more reliable and if possible multilingual methods and procedures that will generate the necessary datasets in a short time and in good quality. This of course goes hand in hand with fundamental research on language models and in general on Deep Learning and hybrid approaches, since progress there can change the need for data in volume, annotation and other aspects. While recently presented results indicate that novel approaches could indeed be applied to address some of the challenges related to the creation of models for low-resourced languages, the scope of their application and inherent limitations are still the subject of ongoing research (Backfried et al., 2022). With regards to accessibility of data the movement of FAIR Data and Principles,[94] that has been established in the scientific community and is now spreading also to other data communities, has become a de facto standard concept. The FAIR principles are: Findability, Accessibility, Interoperability, and Reuse of digital assets. The EU's Data Spaces represent a positive step towards addressing these data availability issues.

Data Interoperability Compounding the issue of Data Availability, Data interoperability is another important factor in regard to data acquisition, sharing and efficient use. There are countless standards regarding data interoperability in place worldwide by several standardisation bodies and in several industry domains. However, this diversity of data interoperability standards presents a problem as there is only little mapping between such standards and approaches. Standards also often are developed in research projects without the involvement of the end users and thereby sometimes lack implementation experience. Moreover, there are many data silos in place that are not connected nor interoperable. Therefore, data acquired from different sources need to be continually tested, evaluated, improved in quality and finally integrated with huge effort, which makes working solutions based on such data extremely costly (Kaltenboeck et al., 2022).

Data and Ethics With the rise of artificial intelligence (AI) and machine learning, as well as the overall movement of data collection and processing, the component of: data and ethics becomes more and more important. Neural language models operate as black boxes that are hard to interpret. This lack of transparency makes it difficult to build trust between human users and system decisions. Lack of explanation abilities is a major obstacle to bring such technology in domains where regulation demands systems to justify every decision. Furthermore, language models face ethical challenges including gender, racial and ethnical biases that are learnt from biases present in the data the models are trained on, thus perpetuating social stereotypes (e.g. as found in MT (Vanmassenhove and Way, 2018; Vanmassenhove et al., 2019)). These biases replicate regrettable patterns of socio-economic domination that

[94] https://www.go-fair.org/fair-principles/

are conveyed through language, since these biases are present in the training data and are then amplified by models which tend to choose more frequent patterns and discard rare ones. In the future, ethical and fair LT should not further propagate notions of inequality, but rather foster an inclusive society based on acceptance and respect: in future models, those biases should be removed altogether, to ensure that the language produced by such systems does not reinforce and propagate inequality and exclusion. Future research avenues include exploring ways to achieve this through the examination of training data, identifying biased parts or gaps, and enriching the data by providing alternatives, or by replacing them altogether. Modifying models could reduce biases, too, for example by introducing weights for probabilities of words related to bias. In terms of ethically and responsibly managing language data, issues relating to data privacy, security, and the processing and protection of personal identifiable information (PII) need to be taken into account, while seeking a balance whereby laws (such as the current GDPR restrictions) do not hamper data usage and technology development.

Knowledge Graphs Knowledge Graphs provide powerful mechanisms and principles to interlink and enrich data in a high quality manner. Thereby Knowledge Graphs can build a powerful and relatively easy to maintain network of interlinked data – including and combining structured, semi-structured and unstructured data – that can be seen as a crucial data infrastructure element to develop future Language Technology Solutions. As such solutions require not only a single underlying dataset but a wide range of meaningful and contextualised data. In addition the integrated data models inside of Knowledge Graphs (taxonomies, vocabularies and/or ontologies) allow the training of algorithms for Language Technology solutions with higher precision and less training data (Kaltenboeck et al., 2022).

AI is often not precise enough in regard to simple challenges that require common sense knowledge or context and meaning (semantics) and thereby "deep language understanding" regarding language technologies. Knowledge Graphs are playing a large part in new approaches that combine the two main fields of AI, namely: statistical AI (machine learning & algorithms) and symbolic AI (models like ontologies, knowledge bases for common sense knowledge, cultural resources). As such, the term Semantic AI has been on increasing used in recent times (Kaltenboeck et al., 2022). Further research in this direction will help to lead to Deep NLU and a move away from pure reliance on large datasets.

5.1.6 Infrastructure-related priority research theme

Hardware infrastructures are required to accommodate for the required computation power and storage of Deep Neural Networks. While in North America and Asia public and private resources can be allocated to only a limited number of languages, to effectively honour the well-entrenched commitment to promote multilingualism in Europe resources must be distributed across a large number of official and unofficial EU languages, so that the respective language communities are treated fairly. As a result, the scale at which European research can be conducted is limited in comparison. There is also an uneven distribution of resources across countries, regions and languages (Aldabe et al., 2021b). Considering the massive infrastructure that is required to train very large state-of-the-art LT systems, Europe starts with a systemic handicap. Europe's strong foundation in research and innovation can compensate for the disadvantage European organisations have with respect to infrastructure, provided that a concerted effort is undertaken in researching the development of new hardware platforms and respective AI training paradigms. At the same time, the hardware on which LT runs must be scaled down. Several approaches to replace GPU-based computing, or at least to make it more power-efficient, are already under investigation. By ensuring that the capabilities of the hardware are aligned with the needs of ML training and inference mod-

els, smaller models would be easier to integrate and use on any device and also be greener by requiring fewer resources, since training neural models is resource-intensive and has a heavy carbon footprint (Strubell et al., 2019a). The EU has the opportunity to be a pioneer in developing such LT models by focusing also on efficiency both in terms of hardware and software. This would not only have positive environmental consequences, but it will also level the paying field for smaller and not well-resourced institutions and companies.

Data and Knowledge Infrastructures In addition to hardware infrastructures, there is also a clear need for a comprehensive and interconnected data infrastructure that needs to be put in place to achieve the specified objectives. To fill the identified gaps in data, language resources, and Knowledge Graphs we recommend and suggest a future path for Europe towards comprehensive and interlinked data infrastructures. These infrastructures have to provide interoperability out-of-the-box by following harmonised and well-proven standards, regarding (i) data (semantic data) interoperability as well as (ii) services and (iii) innovative metadata and data management tools that are available along all steps of the data life cycle. Metadata, data, data-driven services and data-driven tools to be easily docked into these data infrastructures, with the necessary and huge efforts in data cleaning and data integration done only to newly acquired resources (Kaltenboeck et al., 2022). The goal is a federated network and infrastructure of interlinked data spaces for language technology. Existing data spaces as well as newly developed ones should be integrated, where appropriate and possible. Data driven services are provided and can be used along end users requirements. Integrated crowdsourcing and/or citizen science mechanisms allow human-machine interaction to foster data acquisition, cleaning and enrichment (e. g., annotation, classification, quality validation and repair, domain specific model creation, etc.). Raw data can be loaded into available tools to train algorithms or create memories and/or (language) models for specific use cases, but also existing algorithms, models or vocabularies are available and can be easily loaded and re-used to avoid unnecessary energy consumption / computing power to foster the idea of energy efficient data management. In addition, high importance needs to be put on privacy protection (related to personal identifiable information (PII) and beyond), the avoidance of bias (for example on gender), and on data sovereignty. The approach of such data infrastructures require working and sustainable business models that allow data trading, sharing and collaboration. Well targeted publicly funded/supported programmes and activities in the area of data literacy are required from early education onward, to ensure that sufficient human resources in the field are available in the future. In addition an action plan for the collection and the development of data and language resources that are relevant for language technology, as well as for Knowledge Graphs is needed to ensure the availability of sufficient data in the EU languages, as well as in dialects and important non-EU languages. The recommendation for this is to look into three areas: (i) Language Equality Action Plan by means of targeted national and European funding along a matrix of relevant resources and languages, combined with (ii) more measures in the fields of crowdsourcing and citizen science, and (iii) the development of functioning data related business models; beside technology, interoperability or data related attributes there must be a strong focus established on applying all these mechanisms and methodologies to the widest range of languages possible, at least to EU languages but also local and regional dialects of these languages, as well as to non-EU languages that are widespread across Europe. Without such data and language resources in place, a digital language equality cannot be reached (Kaltenboeck et al., 2022). Such data must be easily accessible with fair conditions and costs in a clearly specified legal environment providing transparent rules and regulations. For the European research community to foster innovations in the field, for the industry to successfully compete in a global market, and thereby for the European citizens and its society, that is constantly growing in regard to its diversity and a wide and increasing variety of languages. Data, language

resources, and Knowledge Graphs are thereby a crucial factor on our way to digital European Language Equality, contributing in turn to equal opportunity across all areas of life if European citizens.

5.2 Impact of the European Language Equality Programme

A large-scale, long-term funding activity is needed to push Europe into the leadership position in the field of DLE, LT and Language-Centric AI, and also to secure its leading position for many years to come. The societal and commercial impact of the technologies to be produced by a future ELE programme cannot be overstated: technologies for transparent written and spoken human-human and human-machine communication that are able to cross language barriers and also cultural barriers, as well as technologies that assist organisations and citizens in high-level cognitive tasks, i.e., accessing, translating and making sense of information, obtaining knowledge, communicating with other humans or machines, eventually independently of particular languages or cultures.

Currently, LT as one of the three core application areas within AI together with Vision and Robotics because LT is one of the most important AI application areas with a fast growing societal and economic impact.[95] LT applications such as speech recognition, speech synthesis, textual analysis and machine translation are actually used by hundreds of millions of users on a daily basis. As reflected in the European, national and regional AI and LT strategies both inside and outside Europe, LT is outlined as one of the most relevant technologies for society.

As described in chapter 1.5, this relevance is reflected in recent economic developments. Funding for LT start-ups is booming reflected by the financial support of over USD 1 billion for companies that offer solutions using NLP in 2021.[96] The forecasts done by various consulting firms contain optimistic figures representing an enormous growth of the global LT market based on a significant rise of NLP applications and unstructured digital data. One of these reports from 2021 estimates the growth from USD 20.98 billion in 2021 to USD 127.26 billion in 2028 at a CAGR of 29.4% in the forecasted period .[97] This not only means the need of high qualified new employees on the LT sector but also a very large growth of productivity and efficiency in the whole economy, from industry to administration.

Furthermore, current LT based on Transformers is largely influencing and improving other research areas including text-to-picture generation,[98] genetics,[99], or programming code understanding and generation.[100]

In addition, current multilingual LT with models that cover hundreds of languages opens up the way to leverage NLP techniques for less technologically supported languages, thus promoting DLE in Europe by 2030. The ELE Programme will also consider the gender dimension in the design and impact of LT.

The ELE programme is exactly the large scale of effort that will accelerate the developments and advance the state of the art that will make it possible to join forces that have so far never been joined. This will make it possible to address all European and other relevant languages, all cultures with their particular background and framing of the world, all relevant scientific fields, and all stakeholders by means of a representative number of use cases. This initiative around a giant pool of shared data sets, open evaluations, open competitions, shared tasks, standardisation efforts, etc. in the literal sense of Open Science will

[95] https://oecd.ai/en/classification
[96] https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/?sh=429aff902b14, https://towardsdatascience.com/nlp-how-to-spend-a-billion-dollars-e0dcdf82ea9f
[97] https://www.analyticsinsight.net/the-global-nlp-market-is-predicted-to-reach-us127-26-billion-by-2028/
[98] https://openai.com/dall-e-2/
[99] https://data.solita.fi/reading-the-genomic-language-of-dna-using-neural-networks/
[100] https://www.deepmind.com/blog/competitive-programming-with-alphacode

have an impact in terms of interoperability, development costs, quality and, thus, uptake of the truly game-changing technologies developed in the future programme. Through the ELE Programme, Europe will reclaim scientific and industrial leadership from the US-based monopolists currently dominating the field. The emerging economies in Asia, especially China, have recognised the importance of LT and are investing heavily to bring Chinese companies into leadership positions. Europe, as the prime example for a multilingual society, should see this as motivation to take the leading role that its multilingual nature mandates. The ELE programme may strengthen European LT sovereignty and enhance well-being inside and outside Europe.

6 A Shared European Programme for Language Technology and Digital Language Equality in Europe by 2030: Recommendations

6.1 Overview and Main Concept

The plan and vision described in this document is not only compatible with current EU policy, needs and demands, it is also mission-critical addressing them. Missing investment in the underdeveloped areas of LT and language-centric AI will result in the digital extinction of languages and only global languages spoken by a high number of speakers will prevail. Although the overall EU LT community is quite important, the global LT/NLP market will be dominated by the US and few Asian countries, while the European LT community will be pushed aside.

The main basic requirement of the future ELE Programme is a collaboration between the EU/EC and all participating countries and regions. Moreover, funding and further investment is needed on all levels. Funding on the level of the EU should enable overarching coordination and EU-wide technological infrastructure. It should cover the topics which require pan-European coordination such as shared tasks, protocols, multilingual dataset creation, etc. Increased coordination on European level is needed because language communities are still too fragmented and small. Further effort should be invested into the establishment of the adequate policy-making, distributed research infrastructures and technological platforms like ELG, with flexible access to sufficient HPC facilities. Additionally, national and regional funding should complement the European funding with regard to language-specific research and development. The implementation of these aspects were described, among others, in the ELE language reports.

This section breaks down how concrete recommendations for such a shared programme should look like. First, we outline the possible cornerstones for suitable infrastructure and policy recommendations, as well as ideas for the realisation of a governance model. Second, we revise the technology and data recommendations suggested by the ELE consortium, which are closely related to the ones discusses in the *Language equality in the digital age* resolution (European Parliament, 2018).

Further, research recommendations are considered ground-breaking and game-changing by the LT community. Over the last decade, the community has developed a clear vision of the work needed in the different areas of LT. This vision has been outlined in various strategic research and innovation agendas. The European Parliament has also acted on these ideas. In the last year, the ELE consortium has further investigated these new directions of research.

The need to refocus and massively strengthen European LT/NLP research through a large-scale initiative as a shared, collaborative pan-European effort between EU and participating countries and regions (ELE Programme) has to be agreed upon by all involved parties. Such an endeavour should further increase the participation between research centres, academia,

enterprises (particularly SMEs and start-ups), and other relevant stakeholders. As LT is aggregated and applied to more complex settings, inter-disciplinary research and activities are becoming more relevant in order to further boost developments and allow synergies to become apparent. To achieve *DNLU*, we need to finance and investigate fields such as cognitive, symbolic and pattern-based AI further.

Funding programmes should boost pan-European long-term basic research as well as knowledge and technology transfer between research labs and industry. Frequently mentioned areas and tasks for basic and applied research where further investigation is needed include language data collection (text, dialog, vision, sign language and other forms of interactions), speech analysis, AI, human-computer interaction, machine learning, robotics, natural language understanding and processing tasks such as machine reading, text analysis, machine translation, chatbots, virtual assistants or summarisation.

Further, we are outlining concrete implementation recommendations.

6.2 Policy Recommendations

- To reinforce European leadership in LT by establishing the ELE programme as a large-scale, long-term coordinated funding programme for research, development, innovation and education with the societal goal of digital language equality and the scientific goal of deep natural language understanding.
- To ensure comprehensive EU-level legal protection for the more than 60 regional and minority languages.
- To empower recognition of the collective rights of national and linguistic minorities in the digital world (including sign languages)
- To encourage mother-tongue teaching for speakers of official and non-official languages of the EU.
- To safeguard sufficient funding to support the new technological approaches, based on increased computational power and better access to sizeable amounts of data.
- To develop specific programmes within current funding schemes, especially Horizon Europe and Digital Europe (including the Recovery Plan for Europe), to boost long-term basic research as well as knowledge and technology transfer between countries and regions, and between academia and industry.
- To define and develop a BLARK-like[101] minimum set of language resources and capacities that all European languages should possess.
- To develop common policy actions and clear protocols for language data sharing by public administration at all levels. Language data should be included as a high-value data category in the Open Data Directive (2019/1024/EU).
- To develop clear and robust protocols to ensure flexible access to sufficient GPU-based HPC infrastructure and robust protocols to process sensible data.
- To enable and empower European SMEs and startups to easily access and use LT in order to grow their businesses online independent of language barriers.
- To create the necessary appealing conditions to attract and retain qualified and diverse international LT personnel in Europe.
- To ensure mechanisms to achieve European LT sovereignty.

[101] http://www.blark.org

6.3 Governance Model

- To structure the ELE Programme as a shared, collaborative and coordinated programme between the EU and participating countries and regions.
- To allocate the area of multilingualism, linguistic diversity and language technology to the portfolio of a EU Commissioner.
- To spark a large lobby for EU regional and minority languages (RML).
- To create a pan-European network of research centers to facilitate the coordination of the ELE programme at all levels.
- To promote a distributed centre for linguistic diversity that will strengthen awareness of the importance of lesser-used, regional and minority languages.
- To design and apply new forms of research funding and organisation to ease the transition from application-oriented basic research to commercially focused technology.
- To construct a multilingual LT benchmark, a European "SuperGLUE"-style shared benchmark, that tracks progress.
- To strongly encourage all EC-funded projects to have a language diversity plan and to include direct or associated partners from a less-widely spoken language.
- To facilitate EU Member States' acquisition of LT for their local industries without depending on non-European technology providers.

6.4 Technology and Data Recommendations

- To develop high-performance applications (in terms of speed and quality) for all languages that respect safety, security and privacy.
- To address the lack of available data and define the minimum of language resources and capacities that all European languages should possess.
- To add more focus on systematic language data collection (text, dialogue, multimodal) and exploit automatic data generation (synthetic data), crowd-sourcing and translation of data.
- To ensure efficient adaptations to applications, both in terms of language, domain, efficiency, power consumption, ease of maintenance, and quality assurance.
- To develop methods to overcome the unequal data availability, by focusing on, e. g., annotation transfer, multilingual models preserving quality, few-shot or zero-shot learning.
- To unleash the power of public sector data, data from broadcasters, social media, publishers etc.
- To enforce open ecosystems, open source, open access, open standards and interoperability.
- To focus on research in data bias for strengthening inclusiveness and accessibility.
- To focus upon green LT with a small compute and carbon footprint (e.g., model compression). Green LT (i. e. technologies with low-demand computational footprint).

- To foster publicly available resources that facilitate innovation and research for both commercial and non-commercial actors.
- To develop large open-source language models that work for all EU languages, optimised in in terms of compute time and cost.
- To develop new methodologies for transfer and adaptation of resources and technologies to other domains and languages.
- To define the minimum language resources that all European languages should possess in order to prevent digital extinction.
- To support the coordination between research and industry to enhance the digital possibilities for language translation and open access to the data required for technological advancement.
- To encourage administrations at all levels should improve access to online services and information in different languages.

6.5 Infrastructure Recommendations

- To strengthen existing and create new research infrastructures (RIs) and LT platforms that support research and development activities, including collaboration, knowledge sharing, and open access to data and technologies.
- To ensure sufficient operational capacity, especially for large language models.
- To fill the identified gaps in data,language resources, and knowledge graphs create a future path for Europe towards comprehensive and interlinked data infrastructures.
- The technology vision of an integrated and interoperable data infrastructure shall follow the idea of a Semantic Data Fabric including rich semantics, and thereby context and meaning as well as dynamic and augmented metadata and data management.
- To ensure flexible access to GPU-based HPC facilities and a more suitable computing infrastructure.
- To create an European network of centres of excellence in LT to increase industry visibility, design national research agendas and employ a European Data Strategy.

6.6 Research Recommendations

6.6.1 Recommendations for all LT research areas

- To gather and make available the necessary critical mass of resources in terms of data, computing facilities, and expertise from pan-European LT research labs and centres, with the support from the EC as well as national and regional administrations.
- To create sufficient multilingual and multi-modal data of quality (responsible, legal, diverse, unbiased, ethical, representative, etc.), in all European languages and domains (media, health, legal, education, etc.).
- To provide flexible access to HPC facilities in the form of clusters of high capacity GPUs for LT research and industry. HPC facilities should provide clear and robust protocols to process sensitive data.

- To develop better benchmarks and datasets (ethical, responsible, legal, etc.) for all languages, domains, tasks and modalities.
- To combine interactive LT (conversational AI) with text, knowledge, and multimedia technologies for a new generation of applications that can address the deeper questions of communication, common sense and reasoning.
- To encourage responsible, green, trustworthy, unbiased, inclusive, non-discriminatory LT/AI, making interpretability and explainability of AI models a priority.
- To develop further the areas of Responsible AI and Explainable AI by combining of statistical and symbolic AI in multilingual environments to provide AI-based applications that bring accurate results and benefits for research, industry, and society.
- To focus on methods and learning architectures to overcome the highly unequal data availability, such as annotation transfer, synthetic data and their proper use in machine learning, multilingual models preserving quality and coverage and few-shot or zero-shot learning.
- To focus on Green LT and investigate new efficient methods to extend, reuse and adapt existing pre-trained language models or develop new ones with much reduced carbon footprint.
- To develop language and culture-specific technologies that cover more linguistic phenomena and text types, focusing on accessibility, through sign language, avatar technology etc.

6.6.2 Machine Translation

- To develop direct and near-real-time speech-to-speech MT and adaptive MT, where the system learns from linguists' input.
- To develop low-resource MT, by deepening research on embedding projection and structural organisation of embeddings to apprehend how structurally different languages and their respective embedding spaces can be mapped on to one another.
- To provide transparency of AI models with regard to accuracy and fairness.
- To move towards context-aware methodologies that goes beyond text data and include images, videos, tables, etc. by developing multimodal MT systems.
- To reframe MT, and NLP in general, as a quantum computing problem.

6.6.3 Speech Processing

- To enhance speech resources and create acoustic models to cover a wide variety of languages, including non-standard varieties and dialects.
- To develop good, natural synthetic voices, allowing users to obtain content in their spoken languages.
- To improve context modeling to handle the translation across larger volumes of text.
- To improve the handling of audio conditions currently perceived as difficult (e. g., multiple simultaneous speakers in noisy environments speaking spontaneously and highly emotionally in a mix of languages).

- To support research in the direction of combining speech, NLU and NLP with other modalities, such as image and vision.
- To address privacy and security threats in areas of speech synthesis, voice cloning and speaker recognition.

6.6.4 Text Analytics and Natural Language Understanding

- To increase the adoption of approaches based on self-supervised, zero-shot, and few-shot learning.
- To support research in NLU which integrates speech, NLP, and contextual information as well as additional modes of perception.
- To strengthen basic research in neurosymbolic approaches to NLP/NLU, including grounding and the use of human-understandable databases and sources.
- To create large open-access language models for all European languages (for fine-tuning and downstream tasks), datasets (for training and testing), multilingual models, models that include symbolic knowledge, and models that include discourse features.
- To strengthen progress in reinforcement-based learning, novel dialogue management strategies, and situation-aware natural language generations.
- To strengthen interdisciplinary research and enable better modeling of multimodal environment.

6.7 Implementation Recommendations

- To structure the 9 year long ELE Programme into 3 phases of 3 years each.
- To facilitate discussions between the EU/EC and participating countries to define needs and goals as well as the financial setup.
- To encourage participating countries to invest into the development of LLMs, data sets, technologies, tools for their own languages.
- To have the EU establish binding legislation to encourage or ensure participation.
- To have the EU invest into pan-European coordination of all language-specific projects and initiatives, support mechanisms, infrastructures, data procedures, cross-cutting projects etc. and provide flex funds for bootstrapping poorly supported languages.
- To structure the ELE Programme into 6 themes covering: Language Modelling, Data and Knowledge, Machine Translation, Text Understanding, Speech and Infrastructure. To support each theme by coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs).

7 Roadmap towards Digital Language Equality in Europe by 2030

7.1 Main Components

Human language technologies have the potential to overcome the linguistic divide in the digital sphere. However, we need to define actions, tools, processes and actors that need to be

involved. The goal of this SRIA is to lay out a roadmap with concrete steps for the implementation that carry tangible and measurable outputs and to obtain broad endorsement by the relevant stakeholders.

The main scientific goal of the ELE Programme is *Deep Natural Language Understanding in Europe by 2030* (DNLU). This will increase efficiency by sharing knowledge, infrastructures and resources, with a view to developing innovative technologies and services, in order to achieve the next scientific breakthrough in this area and help reduce the technology gap between European languages with the (interdisciplinary) collaboration of research centres, academic experts, enterprises and other relevant stakeholders. Crucially, such a long-term ELE Programme must involve significantly intensified coordination between the European LT research and the industry.

The main societal and economical goal of the ELE Programme should be *digital language equality in Europe in 2030*. The focus is on language equality and the provisioning of technologies, services and resources outside the often-preferred languages to achieve technological sovereignty in this crucial application area. For minority and lesser spoken languages, we need to find a (technological) way to consider DNLU within a common approach, to create synergies and increase efficiency of the solutions and their design and development. To narrow the digital divide, there is a pressing urgency for novel techniques that would bring less-resourced languages to a level comparable to state-of-the-art results for resource-rich languages. This includes the leveraging of multimodal and multilingual resources to support the development of applications for languages and varieties with scarce resources.

This roadmap towards Digital Language Equality in Europe by 2030 provides a path and the means to ensure that the two goals outlined above are met. To tackle this challenge, the ELE Programme combines the following six themes.

Language Modelling This theme includes research, development and deployment activities regarding LLMs, especially multilingual and multimodal LLMs that include text, speech, image, video etc. Time and resources need to be invested for experiments, new approaches, shared tasks etc. For novel research approaches we need to combine national projects and data sets with international consortia. With regard to innovation and deployment, LLMs will be applied in industrial sectors and use cases.

Data and Knowledge The Data and Knowledge theme is focused on the collection, production, annotation, curation, quality assessment, standardisation etc. of text data, spoken data, video data, and other multimodal data.

Machine Translation The MT theme is focused on improving the automated translation from one natural language into another (including sign languages). While Europe has a strong foundation in this field, research needs to combine novel, groundbreaking approaches with results of the Data and Knowledge as well as Language Modelling themes (see above). The results need to be applied in different industrial sectors and use cases. Deployment needs to be fast, agile and driven by excellent teams.

Text Understanding The Text Understanding theme aims to improve the identification and labelling of linguistic information underlying any natural language text (or other modalities). This requires exploring new strands of research and building on synergies of the other themes. An equally important aspect is applicability in the industry.

Speech The Speech theme addresses one big challenge of the European LT community, i. e., the shift from broad text to broad speech or multimodal processing (including corresponding research towards grounding). While progress in the area of speech applications has been made in the last decade, we also need novel research paradigms. This theme will benefit from the themes Data and Knowledge as well as Language Modelling. The development of relevant industry applications is another goal.

Infrastructure The Infrastructure theme involves the extension and maintenance of platforms such as European Language Grid (ELG). ELG has the potential of functioning as one of the primary platforms to support the activities of the ELE Programme. Moreover, ELG will be the sharepoint for best practises and the development of bridges to other relevant platforms. New features and functionalities need be be implemented for a higher adaptability. Other important factors are the provisioning of GPUs and of standardisation.

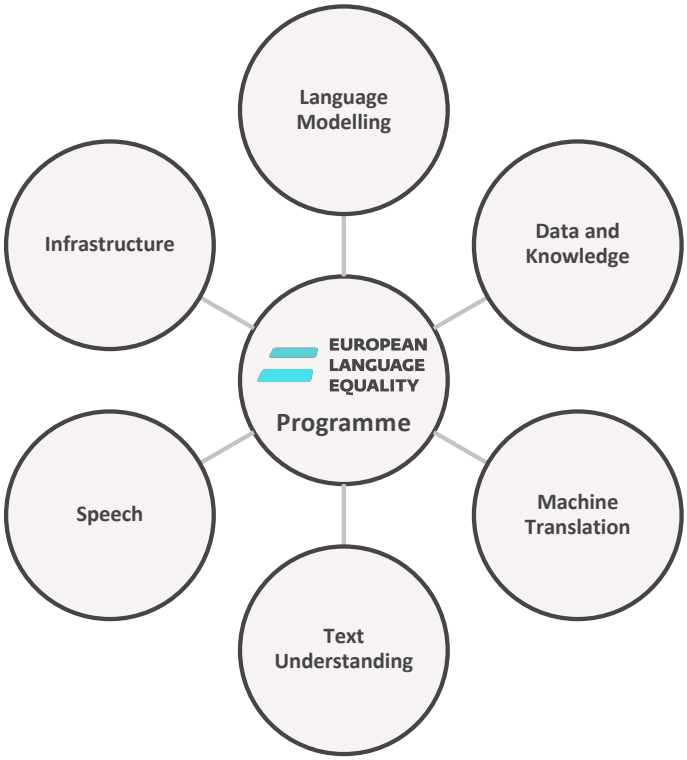

Figure 11: ELE Programme – Themes

7.2 Actions, Budget, Timeline, Collaborations

The European Parliament Resolution "Language equality in the digital age" (September 2018) strongly encourages to "establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, [...] tailored specifically to Europe's needs and demands".

As a direct response, the ELE project has developed an outline of necessary actions. These have been informed by 66 project reports (2400+ pages with condensed findings). A total of 92 languages have been taken into account. We have included voices from research, industry and civil society. For the research sector we have compiled over 30 reports on the situation of individual languages. In addition, we collected input through various surveys and more than 60 expert interviews. To cover the industry angle, our SME partners produced four technical deep dives and collected feedback in a number of surveys for further information. The civil society was represented by the European citizen survey with approx. 20,000 responses.

We foresee an ELE Programme of nine years. This period will be divided into three phases of three years each. The project shall be structured along different themes (Figure 11).

7.2.1 Actions

We foresee different types of projects, implemented using the different typical EC project types: coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs).

Coordination and Support Actions (CSAs) are needed to support research activities and policies (networking, exchange, access to research infrastructures, studies, conferences, etc.). The ELE Programme envisages three CSAs for the overall programme coordination. This includes, among others, the maintenance of the ELE principles, quality assurance approaches, shared tasks etc.

Additional CSAs are needed for the themes *Data and Knowledge* as well as *Language Modelling* as these are fundamental for all other themes as well. Another CSA is needed for infrastructure support.

Research and Innovations Actions (RIA) are collaborative projects funding research activities that allow the exploration of new technologies, new methods, new products, or improvements to existing ones.

Research is an important basis for European Language Technology and Digital Language Equality. Over the last decade, the community has developed a clear vision of the work needed in the different areas of LT. To achieve deep Natural Language Understanding (DNLU), we need to finance and further investigate the fields of language modelling, machine translation, text understanding and speech.

Innovations Actions (IAs) consist of activities directly aiming at producing improved products, processes or services. They may include prototyping, testing, demonstrating, piloting, large-scale product validation and market replication.

	Type	Number
ELE Programme – overall coordination	CSA	3
Theme Data and Knowledge – coordination	CSA	3
Theme Language Modelling – coordination	CSA	3
Theme Language Modelling – research	RIA	15
Theme Language Modelling – innovation and deployment	IA	15
Theme Machine Translation – research	RIA	12
Theme Machine Translation – innovation and deployment	IA	12
Theme Text Understanding – research	RIA	12
Theme Text Understanding – innovation and deployment	IA	12
Theme Speech – research	RIA	12
Theme Speech – innovation and deployment	IA	12
Theme Infrastructure – support	CSA	3

Table 4: ELE Programme – Different types and number of projects foreseen

7.2.2 Budget

As a shared programme between the EU and the participating countries, the final financial set up needs to be discussed between all involved parties. For the EU part of the budget, we suggest the following breakdown (divided by themes).

ELE Programme (overall coordination)	60M€
Data and Knowledge Theme	45M€
Language Modelling Theme	195M€
Machine Translation Theme	120M€
Text Understanding Theme	120M€
Speech Theme	120M€
Infrastructure Theme	30M€
Total	**690M€**
Flexible funds	150M€

Table 5: ELE Programme – Budget breakdown (EU)

In addition to the total sum of 690M€, we consider another 150M€ as *flexible funds* for languages with fragmentary, weak or no technical support since we anticipate that a number of participating countries will require complementary funding from the European Union. A more detailed breakdown of the different themes with their associated project types and runtime is shown in Table 7.

Investments needed on the individual language level are extremely difficult to predict. These national/regional investments are mostly complementary to the EU/EC funding.

We suggest to group the languages into three clusters (see Table 6). However, it needs to be decided which other factors (number of speakers etc.) could play into the clustering before final numbers can be calculated.

Languages with *weak or no support*	40-50M€ each
Languages with *fragmentary support*	30-40M€ each
Languages with *moderate support*	20-30M€ each

Table 6: ELE Programme – Estimated investments required by language

This language-specific funding is foreseen to be provided by the respective participating country or countries. However, the EU should help bootstrap the development of technologies for languages that are not doing well digitally, using the suggested flexible funds. More precise estimates can be provided in early 2023.

7.2.3 Timeline

The ELE Programme is foreseen to have a runtime of nine years, that are divided into three phases of three years each (Table 7).

Phase 1: 2024-2026 Phase 1 lays a strong basic foundation for the overall project. All projects start in Phase 1, except for the Innovation Actions.

Phase 2: 2027-2029 Phase 2 has a strong focus on the Research and Innovation Actions while continuing with the Coordination Actions.

Phase 3: 2030-2032 Phase 3 continues the Coordination Actions and kicks off the Innovation Actions in 2031.

7.2.4 Collaborations

The ELE Programme complements existing related initiatives and organisations. It will make use of the services and resources provided by these initiatives. Figure 12 shows an overview of these different stakeholders, grouped into several broader categories:

- Data spaces and data infrastructures
- Research and research data infrastructures
- Various AI initiatives
- AI on demand platform
- High performance computing
- Standardisation

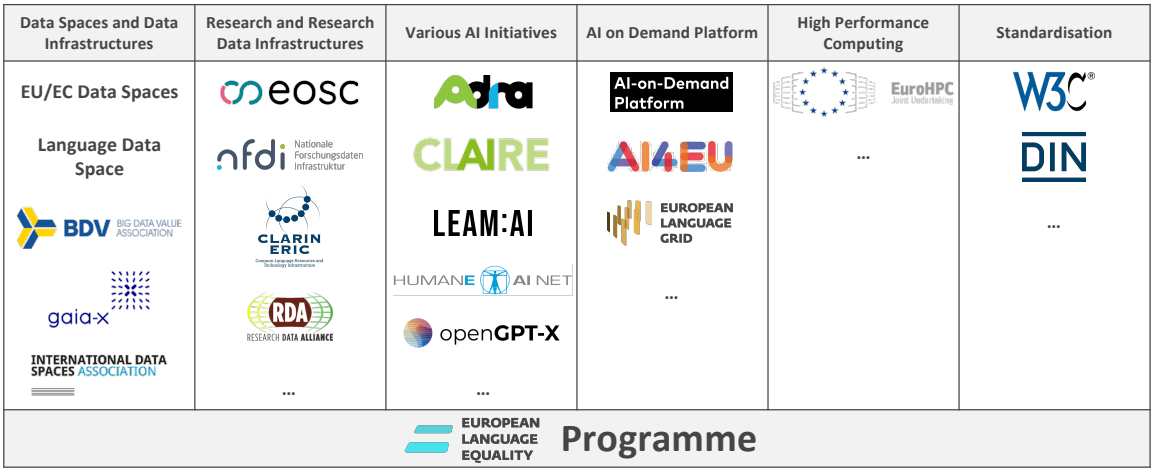

Figure 12: Positioning of the ELE Programme and Foreseen Collaborations

	Type	Num.	Phase 1			Phase 2			Phase 3			Budget	
			2024	2025	2026	2027	2028	2029	2030	2031	2032	Each	Sum
ELE Programme – overall coordination	CSA	3										20M€	60M€
Theme Data and Knowledge – coordination	CSA	3										15M€	45M€
Theme Language Modelling – coordination	CSA	3										15M€	45M€
Theme Language Modelling – research	RIA	15										5M€	75M€
Theme Language Modelling – innovation and deployment	IA	15										5M€	75M€
Theme Machine Translation – research	RIA	12										5M€	60M€
Theme Machine Translation – innovation and deployment	IA	12										5M€	60M€
Theme Text Understanding – research	RIA	12										5M€	60M€
Theme Text Understanding – innovation and deployment	IA	12										5M€	60M€
Theme Speech – research	RIA	12										5M€	60M€
Theme Speech – innovation and deployment	IA	12										5M€	60M€
Theme Infrastructure – support	CSA	3										10M€	30M€
													690M€
Flexible funds for languages with fragmentary, weak or no technological support.													**150M€**

Table 7: ELE Programme – Project types, timeline and budget breakdown (EU)

8 Concluding Remarks

Large-scale studies such as the META-NET White Paper Series (Rehm and Uszkoreit, 2012), the STOA study (STOA, 2017) and the recent ELE language reports (2022) have shown that many languages are in danger of digital extinction because they are not sufficiently supported through Language Technologies. Digital Language Equality is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age (Gaspari et al., 2022a). In alignment with what the Language Technology community has promoted for more than a decade, the European Parliament adopted a resolution on "language equality in the digital age" that suggested initiating a large-scale European LT research, development and innovation programme and to intensify research and funding to achieve deep natural language understanding and also digital language equality (European Parliament, 2018).

Languages are at the heart of every aspect of life. Understanding language is key for building intelligent systems. Over the coming years, AI is expected to transform not only every industry, but also society as a whole. There are diverse trends and megatrends that bear closely on digital technologies. Among others, these include accelerating hyperconnectivity, shifts in the nature of work, increasing digitalisation, new modes of learning, expanding consumerism, novel approaches to politics and governance, changes in healthcare etc. LT and NLP are, by now, considered important driving forces. Current trajectories suggest that LT will play a deciding role in how these unfold.

Language tools and resources have increased and improved since the end of the last century, a process further catalysed by the advent of deep learning and neural networks over the past decade. Indeed, we find ourselves today in the midst of a significant paradigm shift in LT and language-centric AI. This revolution has brought noteworthy advances to the field along with the promise of substantial breakthroughs in the coming years. However, this transformative technology poses problems, from a research advancement, environmental, and ethical perspective. Furthermore, it has also laid bare the acute digital inequality that exists between languages. In fact, many sophisticated NLP systems are unintentionally exacerbating this imbalance due to their reliance on vast quantities of data derived mostly from English-language sources. Other languages lag far behind English in terms of digital presence and even the latter would benefit from greater support. Moreover, the striking asymmetry between official and non-official European languages with respect to available digital resources is very worrisome. The unfortunate truth is that European Language Technology is failing to keep pace with the newfound and rapidly evolving changes in the field.

One need look no further than what is happening today across the diverse topography of state-of-the-art LT and language-centric AI for confirmation of the current linguistic unevenness. The paradox at the heart of recent LT advances is evident in almost every LT discipline. Our ability to reproduce ever better synthetic voices has improved sharply for well-resourced languages, but dependence on large volumes of high-quality recordings effectively undermines attempts to do the same for low-resource languages. Multilingual NMT systems return demonstrably improved results for low- and zero-resource language pairs, but insufficient model capacity continues to haunt transfer learning because large multilingual datasets are required, forcing researchers to rely on English as the best resourced language. A similar language discrepancy is also found in several of the domain sectors: medical corpora, models and knowledge bases suffer from this disparity, as do users of under-resourced languages in education, where access to language-related tools is limited for most smaller language communities.

However, this time of technological transition also represents an opportunity to right the ship; that now is the moment to seek balance between European languages in the digital realm. There are ample reasons for optimism. Although there is more work that can and

must be done, Europe's leading language resource repositories, platforms, libraries, models and benchmarks have begun to make inroads in this regard.

Over the last decade, the community has developed a clear vision of the work needed in the different areas of LT. The ELE project has developed an outline of necessary actions in the form of concrete recommendations. These have been informed by 66 project reports (2400+ pages with condensed findings). A total of 92 languages have been taken into account. We have included voices from research, industry and civil society. For the research sector we have compiled over 30 reports on the situation of individual languages. In addition, we collected input through various surveys and more than 60 expert interviews. To cover the industry angle, our SME partners produced four technical deep dives and collected feedback in a number of surveys for further information. The Civil society was represented by the European citizen survey with approx. 20,000 responses. The DLE metric was developed which calculates for each language the current technological support and the context of the language community in a quantitative manner enabling us to easily compare the support for the languages and to track future developments.

The ELE Programme, specified in the form of this SRIA, will serve as the blueprint for achieving DLE in Europe. While the political and societal goal is indeed reaching full *Digital Language Equality across all European languages* (and, at the same, preventing digital extinction of many of our languages in Europe), the scientific goal envisioned to be reached by 2030 is *Deep Natural Language Understanding*.

Natural Language Understanding is still an open research problem far from being solved since all current approaches have severe limitations. The development of new LT systems would not be possible without sufficient resources (data, experts, compute facilities, etc.). Creation of carefully designed and constructed evaluation benchmarks and annotated data sets for every language and domain of application is needed, to foster technological progress, while encouraging deeper understanding of the mechanisms by which they are achieved. All these efforts will then lead to long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication, to create transparent digital language equality in Europe in all aspects of society, from government to businesses to the citizens.

We foresee an ELE Programme of nine years (2024-2032). This period will be divided into three phases of three years each, combining coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs). The whole community, meaning all relevant scientific and industrial stakeholders from all Member States and Associated Countries need to be involved. Echoing the priority research themes that are currently being discussed, the ELE Programme will tackle the following central themes: Language Modelling, Data and Knowledge, Machine Translation, Text Understanding and Speech.

As a shared programme between the EU and the participating countries, the final financial setup needs to be discussed between all involved parties. We suggest an EU budget of 690M€, plus another 150M€ of flexible funds to help bootstrap the development of technologies for languages with fragmentary, weak or no technical support. This will be supplemented by national and regional funding. With a concerted effort and significant funding, digital language equity will be achieved – for the benefit of all Europeans.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.588.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.468.

Nur Ahmed and Muntasir Wahed. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020. URL https://arxiv.org/abs/2010.15581.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. Report on existing strategic documents and projects in LT/AI, April 2021a. ELE Deliverable D3.1.

Itziar Aldabe, Georg Rehm, and Andy Way. Report on existing strategic documents and projects in lt/ai, 2021b. URL https://european-language-equality.eu/wp-content/uploads/2021/05/ELE___Deliverable_D3_1.pdf.

ALLEA, EASAC, and FEAM. International Sharing of Personal Health Data for Research, 2021. URL https://allea.org/wp-content/uploads/2021/03/International-Health-Data-Transfer_2021_web.pdf.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*, 2021.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1019. URL https://aclanthology.org/P19-1019.

Gerhard Backfried, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz. Deliverable D2.14 Technology Deep Dive – Speech Technologies, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_14__Speech__Technologies.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Aivars Berzins, Khalid Choukri, Maria Giagkou, Andrea Lösch, Helene Mazo, Stelios Piperidis, Mickaël Rigault, Eileen Schnur, Lilli Small, Josef van Genabith, Andrejs Vasiljevs, Andero Adamson, Dimitra Anastasiou, Natassa Avraamides-Haratsi, Núria Bel, Zoltán Bódi, António Branco, Gerhard Budin, Virginijus Dadurkevicius, Stijn de Smeytere, Hrístina Dobreva, Rickard Domeij, Jane Dunne, Kristine Eide, Claudia Foti, Maria Gavriilidou, Thibault Grouas, Normund Gruzitis, Jan Hajic, Barbara Heinisch, Verónique Hoste, Arne Jönsson, Fryni Kakoyianni-Doa, Sabine Kirchmeier, Svetla Koeva,

Lucia Konturová, Jürgen Kotzian, Simon Krek, Gauti Kristmannsson, Kaisamari Kuhmonen, Krister Lindén, Teresa Lynn, Armands Magone, Hélène Mazo, Maite Melero, Laura Mihailescu, Simonetta Montemagni, Micheál Õ Conaire, Jan Odijk, Maciej Ogrodniczuk, Pavel Pecina, Jon Arild Olsen, Bolette Sandford Pedersen, David Perez, Andras Repar, Ayla Rigouts Terryn, Eirikur Rögnvaldsson, Mike Rosner, Nancy Routzouni, Claudia Soria, Alexandra Soska, Donatienne Spiteri, Marko Tadic, Carole Tiberius, Dan Tufis, Andrius Utka, Paolo Vale, Piet van den Berg, Tamás Váradi, Kadri Vare, Andreas Witt, Francois Yvon, Janis Ziedins, and Miroslav Zumrik. *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe - Why Language Data Matters: ELRC White Paper*. ELRC Consortium, 2 edition, 2019a. ISBN 978-3-943853-05-6.

Aivars Berzins, Khalid Choukri, Maria Giagkou, Andrea Lösch, Helene Mazo, Stelios Piperidis, Mickaël Rigault, Eileen Schnur, Lilli Small, Josef van Genabith, Andrejs Vasiljevs, Andero Adamson, Dimitra Anastasiou, Natassa Avraamides-Haratsi, Núria Bel, Zoltán Bódi, António Branco, Gerhard Budin, Virginijus Dadurkevicius, Stijn de Smeytere, Hrístina Dobreva, Rickard Domeij, Jane Dunne, Kristine Eide, Claudia Foti, Maria Gavriilidou, Thibault Grouas, Normund Gruzitis, Jan Hajic, Barbara Heinisch, Verónique Hoste, Arne Jönsson, Fryni Kakoyianni-Doa, Sabine Kirchmeier, Svetla Koeva, Lucia Konturová, Jürgen Kotzian, Simon Krek, Gauti Kristmannsson, Kaisamari Kuhmonen, Krister Lindén, Teresa Lynn, Armands Magone, Maite Melero, Laura Mihailescu, Simonetta Montemagni, Micheál Õ Conaire, Jan Odijk, Maciej Ogrodniczuk, Pavel Pecina, Jon Arild Olsen, Bolette Sandford Pedersen, David Perez, Andras Repar, Ayla Rigouts Terryn, Eirikur Rögnvaldsson, Mike Rosner, Nancy Routzouni, Claudia Soria, Alexandra Soska, Donatienne Spiteri, Marko Tadic, Carole Tiberius, Dan Tufis, Andrius Utka, Paolo Vale, Piet van den Berg, Tamás Váradi, Kadri Vare, Andreas Witt, Francois Yvon, Janis Ziedins, and Miroslav Zumrik. *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe - Why Language Data Matters: ELRC White Paper*. ELRC Consortium, 2 edition, 2019b. ISBN 978-3-943853-05-6.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, 2020.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022. URL https://arxiv.org/abs/2204.06745.

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

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng,

Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

Daan Broeder, David Nathan, Sven Strömqvist, and Remco Van Veenendaal. Building a federation of language resource repositories: the DAM-LR Project and its continuation within CLARIN. In *Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. ELRA, 2008.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Bundesministerium für Wirtschaft und Energie. The Gaia-X Hub Germany. https://www.bmwk.de/Redaktion/EN/Dossier/gaia-x.html, 2020. Accessed: 2022-05-16.

Aivars Bērziņš, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, and Andy Way. Deliverable D2.13 Technology Deep Dive – Machine Translation, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_13__Machine_Translation_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*, 2021. URL https://arxiv.org/abs/2103.12028.

Noam. Chomsky. *Syntactic structures.* The Hague: Mouton., 1957.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Baindoor Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith. Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 99–104, 2019.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12 (ARTICLE):2493–2537, 2011.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*, 2021.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/7137debd45ae4d0ab9aa953017286b20-Abstract.html.

Robert Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

Nicholas Davis and Thomas Philbeck. Understanding the Risk Landscape. In *Global Risks Report 2017*, number 12 in Global Risks Report, pages 43–47. World Economic Forum, 2017. URL https://reports.weforum.org/global-risks-2017/.

Iria de Dios-Flores, Carmen Magarinos, Adina Ioana Vladu, John E Ortega, José Ramom Pichel, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarın, Manuel González González, et al. The nós project: Opening routes for the galician language in the field of language technologies. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 52, 2022.

Franciska de Jong, Bente Maegaard, Darja Fišer, Dieter van Uytvanck, and Andreas Witt. Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3406–3413. European Language Resources Association, May 2020. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.417.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.292. URL https://aclanthology.org/2020.findings-emnlp.292.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. Enhancing transformer for end-to-end speech-to-text translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 21–31, Dublin, Ireland, 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6603.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning, 2021.

Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*, 2021. URL https://arxiv.org/abs/2107.00676.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Gómez-González E and Gomez Gutierrez E. Artificial intelligence in medicine and healthcare: applications, availability and societal impact. Scientific analysis or review, Publications Office of the European Union, Luxembourg, 2020.

European Commission, Joint Research Centre, J Rudkin, L Kimbell, E Stoermer, F Scapolo, and L Vesnic-Alujevic. *The future of government 2030+ : a citizen centric perspective on new government models*. Publications Office, 2019. doi: 10.2760/145751.

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.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. On making reading comprehension more comprehensive. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 105–112, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5815. URL https://aclanthology.org/D19-5815.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. Investigating the impact of gender representation in asr training data: a case study on librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92. Association for Computational Linguistics, 2021.

Federico Gaspari, Andy Way, Jane Dunne, Georg Rehm, Stelios Piperidis, and Maria Giagkou. Deliverable D1.1 Digital Language Equality (preliminary definition), 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/05/ELE_Deliverable_D1_1.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Federico Gaspari, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. Introducing the digital language equality metric: Technological factors. In *Proceedings of The Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 1–12, Marseille, France, June 2022a. European Language Resources Association. URL https://aclanthology.org/2022.tdle-1.1.

Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way. Deliverable D1.3 Digital Language Equality (full specification), 2022b. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, and Valerie Mapelli. The META-SHARE Metadata Schema for the Description of Language Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1090–1097, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

M Goldstein. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech communication*, 16(3):225–244, 1995.

Jose Manuel Gomez-Perez, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, and Teresa Lynn. Deliverable D2.15 Technology Deep Dive – Text Analytics, Text and Data Mining, NLU, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_15__Text_Analytics_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

Annika Grützner-Zahn and Georg Rehm. Introducing the digital language equality metric: Contextual factors. In *Proceedings of The Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 13–26, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.tdle-1.2.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, et al. Pre-trained models: Past, present and future. *AI Open*, 2021.

Erhard Hinrichs and Steven Krauwer. The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1525–1531, 2014.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

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.

Martin Kaltenboeck, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg. Deliverable D2.16 Technology Deep Dive – Data, Language Resources, Knowledge Graphs, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D2_16__Data_and_Knowledge_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Naoyuki Kanda, Xuankai Chang, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. Investigation of end-to-end speaker-attributed asr for continuous multi-talker recordings. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 809–816. IEEE, 2021a.

Naoyuki Kanda, Guoli Ye, Yu Wu, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. Large-scale pre-training of end-to-end multi-talker asr for meeting transcription with single distant microphone. *arXiv preprint arXiv:2103.16776*, 2021b.

Hiroaki Kitano. Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery. *AI Magazine*, 37(1):39–49, Apr. 2016. doi: 10.1609/aimag.v37i1.2642. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2642.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the 6th Conference on Machine Translation (WMT 2021)*, 2021. URL https://arxiv.org/abs/2107.10821. 17pp.

András Kornai. Digital language death. *PloS one*, 8(10):e77056, 2013.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Jennifer Lawlor, Carl Thomas, Andrew T Guhin, Kendra Kenyon, Matthew D Lerner, UCAS Consortium, and Amy Drahota. Suspicious and fraudulent online survey participation: Introducing the real framework. *Methodological Innovations*, 14(3):20597991211050467, 2021. doi: 10.1177/20597991211050467. URL https://doi.org/10.1177/20597991211050467.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.302.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Jason Phang, Ofir Press, et al. What language model to train if you have one million gpu hours? In *ACL Workshop "Challenges & Perspectives in Creating Large Language Models"*, 2022.

Sean M. Leahy, Charlotte Holland, and Francis Ward. The digital frontier: Envisioning future technologies impact on the classroom. *Futures*, 113:102422, 2019. ISSN 0016-3287. doi: https://doi.org/10.1016/j.futures.2019.04.009. URL https://www.sciencedirect.com/science/article/pii/S0016328718304166.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Tal Linzen. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.465. URL https://aclanthology.org/2020.acl-main.465.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. Neural cross-lingual event detection with minimal parallel resources. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 738–748, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1068. URL https://aclanthology.org/D19-1068.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488, May 2021.

LT-Innovate. The lt-innovate innovation agenda, 2016. URL http://www.lt-innovate.org/sites/default/files/2904-LTi_Innovation_Agenda.pdf.

Yinquan Lu, Haonan Lu, Guirong Fu, and Qun Liu. Kelm: Knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs. *arXiv preprint arXiv:2109.04223*, 2021.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *35th AAAI Conference on Artificial Intelligence*, 2021.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.645. URL https://aclanthology.org/2020.acl-main.645.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a. URL https://arxiv.org/abs/1301.3781.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013b. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1008.

George A. Miller. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. URL https://aclanthology.org/H92-1116.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021. URL https://arxiv.org/abs/2110.15943.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, 2022.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, 2019.

John R Pierce and John B Carroll. Language and machines: Computers in translation and linguistics, 1966.

Stelios Piperidis, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi. META-SHARE: One Year After. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1532–1538, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. 12pp.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. URL https://arxiv.org/abs/2102.12092.

ITUT Rec. P. 800.1, mean opinion score (mos) terminology. *International Telecommunication Union, Geneva*, 2006.

Georg Rehm. Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda, November 2017. URL http://cracker-project.eu/sria/. Version 1.0. Unveiled at META-FORUM 2017 in Brussels, Belgium, on November 13/14, 2017. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded project CRACKER.

Georg Rehm and Stefanie Hegele. Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, pages 3282–3289, Miyazaki, Japan, May 2018. 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, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Váradi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, and Sigve Gramstad. An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In Laurette Pretorius, Claudia Soria, and Paola Baroni, editors, *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, pages 30–37, Reykjavik, Iceland, May 2014.

Georg Rehm, Jan Hajic, Josef van Genabith, and Andrejs Vasiļjevs. Fostering the Next Generation of European Language Technology: Recent Developments – Emerging Initiatives – Challenges and Opportunities. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 1586–1592, Portorož, Slovenia, May 2016a. European Language Resources Association (ELRA).

Georg Rehm, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asunción Moreno, Jan Odjik, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Sandford Pedersen, Inguna Skadiņa, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiļjevs, Kadri Vider, and Jolanta Zabarskaite. The Strategic Impact of META-NET on the Regional, National and International Level. *Language Resources and Evaluation Journal*, 50(2):351–374, 2016b. 10.1007/s10579-015-9333-4.

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 2020a. European Language Resources Association (ELRA).

Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon. The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. 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 3315–3325, Marseille, France, 5 2020b. European Language Resources Association (ELRA).

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals. European Language Grid: A Joint Platform for the European Language Technology Community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*, pages 221–230, Kyiv, Ukraine, 4 2021. Association for Computational Linguistics (ACL).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computa-*

tional Linguistics (Volume 1: Long Papers), pages 856–865, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1079. URL https://aclanthology.org/P18-1079.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1621. URL https://aclanthology.org/P19-1621.

Morgane Riviere, Jade Copet, and Gabriel Synnaeve. Asr4real: An extended benchmark for speech models. *arXiv preprint arXiv:2110.08583*, 2021.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54.

Rudolf Rosa, Ondřej Dušek, Tom Kocmi, David Mareček, Tomáš Musil, Patrícia Schmidtová, Dominik Jurko, Ondřej Bojar, Daniel Hrbek, David Košťák, Martina Kinská, Josef Doležal, and Klára Vosecká. Theaitre: Artificial intelligence to write a theatre play. In *Proceedings of AI4Narratives2020 workshop at IJCAI2020*, 2020.

De Nigris S, Craglia M, Nepelski D, Hradec J, Gomez-Gonzales E, Gomez Gutierrez E, Vazquez-Prada Baillet M, Righi R, De Prato G, Lopez Cobo M, Samoili S, and Cardona M. Ai watch : Ai uptake in health and healthcare, 2020. Technical report, Publications Office of the European Union, Luxembourg, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (05):8732–8740, 2020.

Ruslan Salakhutdinov. Deep learning. In Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani, editors, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, page 1973. ACM, 2014. doi: 10.1145/2623330.2630809. URL https://doi.org/10.1145/2623330.2630809.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. URL https://arxiv.org/abs/2110.08207.

Abulhair Saparov and Tom M Mitchell. Towards general natural language understanding with probabilistic worldbuilding. *Transactions of the Association for Computational Linguistics*, 10:325–342, 2022.

Dave Sayers, Rui Sousa-Silva, Sviatlana Höhn, Lule Ahmedi, Kais Allkivi-Metsoja, Dimitra Anastasiou, Lynne Beňuš, Štefan; Bowker, Eliot Bytyçi, Alejandro Catala, Anila Çepani, Sami Chacón-Beltrán, Rubén; Dadi, Fisnik Dalipi, Vladimir Despotovic, Agnieszka Doczekalska, Sebastian Drude, Robert Fort, Karën; Fuchs, Christian Galinski, Christian Galinski, Christian Galinski, Federico Gobbo, Tunga Gungor, Siwen Guo, Klaus Höckner, PetraLea Láncos, Tomer Libal, Tommi Jantunen, Dewi Jones, Blanka Klimova, EminErkan Korkmaz, Mirjam Sepesy Maučec, Miguel Melo, Fanny Meunier, Bettina Migge, Verginica Barbu Mititelu, Arianna Névéol, Aurélie; Rossi, Antonio Pareja-Lora, Aysel Sanchez-Stockhammer, C.; Şahin, Angela Soltan, Claudia Soria, Sarang Shaikh, Marco Turchi, Sule Yildirim Yayilgan, Maximino Bessa, Luciana Cabral, Matt Coler, Chaya Liebeskind, Ilan Kernerman, Rebekah Rousi, and Cynog Prys. The dawn of the human-machine era : A forecast of new and emerging language technologies. Technical report, LITHME project, 2021. URL http://urn.fi/URN:NBN:fi:jyu-202105183003.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.20.

Natalia Simon, Ioannis Markopoulos, Stefan Gindl, Bert Utermark, Martin Kaltenböck, Antragama Ewa Abbas, Hosea Ofe, Montijn van de Ven, Rômy Bergman, Anneke Zuiderwijk, Mark de Reuver, Nora Gras, Antonia Kuster, Julia Jakuzzi, Sebastian Emons, Gerrit Rosam, Michael Fribus, and Alina Brockob. D2.1 'Definition and analysis of the EU and worldwide data market trends and industrial needs for growth', 2021. URL https://www.trusts-data.eu/wp-content/uploads/2021/07/D2.1-Definition-and-analysis-of-the-EU-and-worldwide-data-market-trends-....pdf. The deliverable was written within the project "TRUSTS Trusted Secure Data Sharing Space".

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/.

STOA. Horizon scanning and analysis of techno-scientific trends: Scientific Foresight Study, July 2017. Published: STOA study (PE 603.183), July 2017. Carried out by the Augmented Intelligence Institute (Germany) 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.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019a.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

Support Centre for Data Sharing. Digital Europe Programme explained: the Language Data Space. https://eudatasharing.eu/news/digital-europe-programme-explained-language-data-space, 2022. Accessed: 2022-05-16.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, 2019.

Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf. Language (re)modelling: Towards embodied language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6268–6281, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.559. URL https://aclanthology.org/2020.acl-main.559.

Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020. URL https://arxiv.org/abs/2003.01200.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook ai's wmt21 news translation task submission. In *Proc. of WMT*, 2021.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

Eva Vanmassenhove and Andy Way. SuperNMT: Neural machine translation with semantic supersenses and syntactic supertags. In *Proceedings of ACL 2018, Student Research Workshop*, pages 67–73, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-3010. URL https://aclanthology.org/P18-3010.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6622.

Andrejs Vasiljevs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi. Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem, 2019. DOI 10.2759/142151. A study prepared for the European Commission, DG Communications Networks, Content & Technology by Crosslang, Tilde, ELDA, IDC.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, and Jana Voße. Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 105–110, 2019.

Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, 2021.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. URL https://arxiv.org/abs/2109.01652.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.400. URL https://aclanthology.org/2020.acl-main.400.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.572.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Appendix

A The European Citizen Survey: Supplementary Information

A.1 Survey Questions

1. Please select all the words and terms you are familiar with or that you are able to understand right away: *(Machine Translation, Chatbot, Language Technology, Smart Personal Assistant, Natural Language Understanding, Speech Procesing, Information Retrieval, Language-Centric Artificial Intelligence, Natural Langauge Processing, Conversational Agent)*
2. Where do you live?
3. What is your age? *(18-24, 25-34, 35-44, 45-54, 55-64, 65+)*
4. What is the highest degree or level of school you have completed? If you are currently student, the highest award achieved to date. *(Some High School, High School, Bachelor's Degree, Master's Degree, PhD or higher, Vocational training, Prefer not to say)*
5. What languages do you use in your everyday life (professionally and socially)? Please select as many as apply. *(85 EU-language choice)*
6. Please rate all the types of software applications, apps, tools or devices you use for your language(s). Tools you do not use for your language(s) do not need to be rated.
7. In general, what holds you back from using some of these apps or tools in your languages?
8. Are you aware of any language apps or tools for other languages that you would also like to use for your own languages? (Yes/No)
9. Please select the tools that you currently do not use but would like to use in the future. *(Multiple choice)*
10. What would be the top 3 advantages of improving apps and tools for all languages? Please select the three most important advantages in your opinion. *(See Figure 9 for options)*
11. Do you have any comments you would like to share with us?

A.2 Survey Data Cleaning and Preparation

Collecting survey responses through paid online services can present some known issues that can render results unreliable, at least to some extent (Lawlor et al., 2021). Therefore unsurprisingly, a preliminary analysis of the responses collected revealed some unexpected results. This was particularly true for some languages that received evaluations that diverged from the opinions of ELE language experts. A more in-depth examination was therefore carried out in order to guarantee the integrity of the data that would be subsequently analysed

– the results of which are presented here. Our observation revealed a number of flags indicating unreliable responses, which we filtered from the dataset. As such, a final 20,586 (out of 21,108) responses were analysed. The filtering criteria were as follows:

- 80 responses were removed due to responses coming from duplicate IP addresses, which made it likely that the same person had completed multiple questionnaires. In these cases, only one response from each IP address was retained for data analysis, and all other responses from the same IP address were removed; this data cleaning procedure is reflected in the results presented in Figures 7, 8 and 9.
- 174 responses were removed for inconsistent answers to Question 1. These respondents chose *all* of the terms on the list *including* the "I am not familiar with these terms" option. This clearly revealed an illogical inconsistency and lack of care or sincerity in their response; this data filtering process is reflected in the results presented in Figure 7.
- Question 6 required respondents to rate a number of LT tools for their language. However, we observed inflated scores for a number of languages, in particular for a number of low-resourced languages. The most reliable flag we could use for filtering noisy data at this level were streaks of 4 and 5 stars for languages, for which some of the tools selected do not even exist (e.g. chatbots, screenreaders, automatic subtitling) – deeming the responses unreliable.[102] As such, 581 of these responses were removed. This data cleaning is reflected in the results presented in Figure 8. Table 8 shows the 17 languages to which this applied and the respective number of excluded responses for each of them.

Romanian	116	Croatian	37	Galician	9
Polish	72	Norwegian (Bokmäl)	32	Basque	8
Hungarian	69	Danish	24	Irish	4
Bulgarian	64	Latvian	17	Norwegian (Nynorsk)	4
Slovak	46	Slovenian	16	Catalan	3
Greek	45	Lithuanian	15		

Table 8: Breakdown per language of the excluded responses due to high quality ratings assigned to non-existent LT tools

A.3 Question 6: Calculations Explained

Due to the large size of the dataset and the varying proportion of responses for each language, the final figures presented in Figure 8 are based on the calculation of the median score (purple) and the mode (blue). A median score calculation is less sensitive to outliers than a mean score calculation, and the mode indicates the most frequent rating assigned to a tool for that language. The mode is a useful reference as these are aggregated scores across a number of tools which vary in advancements and reliability across all languages.

Note that tools that were not available or used by a respondent did not receive a score. In these instances, the tool was assigned a rating of zero, as a penalisation for lesser-used tools across all languages. This allowed us to make more accurate inferences about the tools that respondents did actually use and rate. This explains the low scores for languages such as Serbian, Luxembourgish and Icelandic, which either have very few available LTs or low-rated existing LTs.

[102] Following consultation with ELE language informants.

B List of Contributors

Table 9: Organisations which contributed to the SRIA

Organisation	Country
Dublin City University	Ireland
Deutsches Forschungszentrum für künstliche Intelligenz	Germany
Univerzita Karlova	Czech Republic
Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	Greece
Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea	Spain
CROSSLANG NV	Belgium
European Federation of National Institutes for Language	Luxembourg
Réseau européen pour l'égalité des langues	France
European Civil Society Platform for Multilingualism	Denmark
CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	Netherlands
Universiteit Leiden	Netherlands
Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	Germany
Stichting LIBER (Association of European Research Libraries)	Netherlands
Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	Germany
Tilde SIA	Latvia
Evaluations and Language Resources Distribution Agency	France
Expert System Iberia SL	Spain
HENSOLDT Analytics GmbH	Austria
Xcelerator Machine Translations Ltd. (KantanMT)	Ireland
PANGEANIC-B. I. Europa SLU	Spain
Semantic Web Company GmbH	Austria
SIRMA AI EAD (Ontotext)	Bulgaria
SAP SE	Germany
Universität Wien	Austria
Universiteit Antwerpen	Belgium
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	Bulgaria
Sveučilište u Zagrebu Filozofski fakultet	Croatia
Københavns Universitet	Denmark
Tartu Ulikool	Estonia
Helsingin Yliopisto	Finland
Centre National de la Recherche Scientifique	France
Nyelvtudományi Kutatóközpont	Hungary
Stofnun Árna Magnússonar í íslenskum fræðum	Iceland
Fondazione Bruno Kessler	Italia
Latvijas Universitātes Matemātikas un Informātikas institūts	Latvia
Lietuvių Kalbos Institutas	Lithuania
Luxembourg Institute of Science and Technology	Luxembourg
Università ta Malta	Malta
Stichting Instituut voor de Nederlandse Taal	Netherlands
Språkrådet	Norway
Instytut Podstaw Informatyki Polskiej Akademii Nauk	Poland
Universidade de Lisboa, Faculdade de Ciências	Portugal
Institutul de Cercetări Pentru Inteligență Artificială	Romania
University of Cyprus, French and European Studies	Cyprus
Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied	Slovakia

Continued on next page

Table 9 – *Continued from previous page*

Organisation	Country
Institut Jožef Stefan	Slovenia
Centro Nacional de Supercomputación	Spain
Kungliga Tekniska högskolan	Sweden
Universität Zürich	Switzerland
University of Sheffield	United Kingdom
Universidad de Vigo	Spain
Bangor University	United Kingdom
Univerzitet u Beogradu	Serbia
Institut za jezik Univerziteta u Sarajevu	Bosnia and Herzegovina
UiT Noregs arktiske universitet	Norway
Göteborgs universitet	Sweden
Dansk Sprognævn	Denmark
Institutet för språk och folkminnen	Sweden
Nasjonalbiblioteket	Norway
The Language Secretariat of Greenland	Greenland
Mercator	Netherlands

Table 10: Experts consulted

Name	Affiliation	Country
Alexander Baratsits	Cultural Broadcasting Archive	Austria
Marc Berthiaume	European Commission	Belgium
Jörg Bienert	KI Bundesverband e. V.	Germany
Patrick Bunk	Ubermetrics Technologies GmbH	Germany
Leon Derczynski	IT University of Copenhagen	Denmark
Alexander Drechsel	European Parliament	Belgium
Florian Faes	Slator	Switzerland
Jóhanna Vigdís Guðmundsdóttir	Almannarómur	Iceland
Jussi Karlgren	Gavagai and KTH Royal Institute of Technology	Sweden
Peggy van der Kreeft	Deutsche Welle	Germany
Krister Lindén	University of Helsinki	Finland
Arle Lommel	CSA Research	USA
John McCrae	Insight Centre for Data Analytics	Ireland
Arantza Del Pozo	Vicomtech	Spain
Alexander Siebert	Retresco GmbH	Germany
Themos Stafylakis	Omilia Ltd.	Greece
Michael Stormbom	Lingsoft Oy	Finland
Georg Tschare	Sign Time GmbH	Austria
Hans Uszkoreit	DFKI GmbH and Giance GmbH	Germany
Sebastian Wohlrapp	Field 33 GmbH	Germany

Table 11: Organisations represented in the consultation process

Organisation	Country
LT Developers	
4i intelligent insights	Spain
A Data Pro	Bulgaria
Accademia della Crusca	Italy
Adam Mickiewicz University	Poland
AGI - Information Management Consultants	Germany
Ai4Value	Finland
ALAN Analytics s.r.o	Slovakia
AlfaNum	Serbia
Almannarómur / The Voice of the People	Iceland
Amu	Italy
Analyse & Tal	Denmark
Aristotle University of Thessaloniki	Greece
Athena Research Center	Greece
Athens University of Economics and Business	Greece
Audio-Visual Machine Perception Limited	UK
Austrian Research Institute for Artificial Intelligence	Austria
Autonomous University of Barcelona	Spain
Bangor University	UK
Barcelona Supercomputing Center	Spain
Bulgarian Academy of Sciences	Bulgaria
Center for Cultural Protection and Technological Development of Georgian State Languages	Georgia
Center for the Greek Language	Greece
Centre for Aromanian Language and Culture	Bulgaria
Cerence	USA
CERTH	Greece
Charles University	Czech Republic
Ciklopea d.o.o.	Slovenia
CIP4N GmbH	Germany
Cloudwise	Netherlands
CNRS	France
Consiglio Nazionale delle Ricerche	Italy
Convforth SRL	Italy
Cornelistools B.V.	Netherlands
Cyprus University of Technology	Cyprus
Czech Academy of Sciences	Czech Republic
Dalle Molle Institute for Artificial Intelligence	Switzerland
Danish Language Council	Denmark
Darmstadt University of Applied Sciences	Germany
Deloitte	UK
Dublin City University	Ireland
E4 Computer Engineering SpA	Italy
EDIA	Netherlands
emagine GmbH	Germany
EML Speech Technology GmbH	Germany
Ensoul	UK
Entefy	USA
EPFL / Idiap Research Institute	Switzerland
Eurac Research	Italy
Fondazione Bruno Kessler	Italy

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
FORTH	USA
Fraunhofer Gesellschaft	Germany
Free University of Bozen-Bolzano	Italy
Furtwangen University	Germany
Globalese	Germany
Goethe-University Frankfurt	Germany
Grammatek	Iceland
HENSOLDT Analytics	Austria
Heriot-Watt University	UK
HiTZ Basque Center for Language Technology	Spain
Hof University of Applied Sciences	Germany
Human Centered Data Analytics. Centrum Wiskunde & Informatica	Netherlands
Hungarian Research Centre for Linguistics	Hungary
Ilia State University	Georgia
Institute of Philosophy. Czech Academy of Science	Czech Republic
Institute of the Lithuanian Language	Lithuania
Intelartes Sprl	Belgium
Ionian University	Greece
Jožef Stefan Institute	Slovenia
JSC I-Teco	Russia
K Dictionaries - Lexicala	Israel
KantanAI	Ireland
Kempelen Institute of Intelligent Technologies	Slovakia
Kielikone Oy	Finland
KU Leuven	Belgium
LAB University of Applied Sciences	Finland
Laboratoire Hubert Curien	France
Le français des affaires / CCI Paris Ile-de-France	France
Lexical Computing	UK
Lingsoft	Finland
Linköping University	Sweden
LT3. Ghent University	Belgium
Lucid	Netherlands
Lund University Humanities Lab	Sweden
Luxembourg Institute of Science and Technology	Luxembourg
Maastricht University	Netherlands
Macedonian Academy of Sciences and Arts	North Macedonia
magiquo data live s.a	Spain
Masaryk University	Czech Republic
Massey University	New Zealand
Meddal.com	UK
Medical University of Vienna	Austria
Meltwater Group	Norway
Memsource a.s.	Czech Republic
Moravská zemská knihovna v Brně	Czech Republic
Morningsun Technology GmbH	Germany
Mozaika	Bulgaria
Multilingues21. Lda.	Portugal
Národní filmový archiv. Prague	Czech Republic
National and Kapodistrian University of Athens	Greece
National University of Ireland Galway	Ireland
National Centre of Scientific Rsearch “Demokritos”	Greece
Netherlands eScience Center	Netherlands

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
nettle.ai	Slovakia
New York University	USA
Nico van de Water Linguistic Services	Netherlands
Omilia	Cyprus
Pangeanic	Spain
Phonexia s.r.o.	Czech Republic
Polish Academy of Sciences	Poland
Polish-Japanese Academy of Information Technology	Poland
Research Institute for Artificial Intelligence “Mihai Draganescu”, Romanian Academy	Romania
Royal Netherlands Academy of Arts and Sciences	Netherlands
RTL	Germany
Ruhr-Universität Bochum	Germany
Russian Academy of Sciences	Russia
RWS	UK
Samsung Electronics	South Korea
Sberbank	Russia
SciFY PNPC	Greece
Scriptix	Netherlands
SEMLAB	Netherlands
Serbian Academy of Sciences and Arts	Serbia
Sign Time GmbH	Austria
Sinequa	France
Sirma AI (Ontotext)	Bulgaria
Slovak Academy of Sciences	Slovakia
Slovenian Academy of Sciences and Arts	Slovenia
Spanish Society for Natural Language Processing (SEPLN)	Spain
SpeechTech	Czech Republic
Stockholm University	Sweden
Sunda Systems Oy	Finland
Syllabs	France
Talkie.ai	USA
Tallinn University of Technology	Estonia
Technische Universität Dresden	Germany
Text Technology Lab / Goethe University Frankfurt	Germany
The Árni Magnússon Institute for Icelandic Studies	Iceland
The Citizens’ Association for the Promotion of Roma Education “Otaharin”	Bosnia and Herzegovina
The Language Council of Sweden at the Institute for Language and Folklore	Sweden
The MAMA AI. SE	Czech Republic
The National Library of the Czech Republic	Czech Republic
The Welsh Government	UK
Tilburg University	Netherlands
TILDE	Latvia
TMServe	Greece
Toros University	Turkey
Trinity College Dublin	Ireland
Trust Stamp	USA
UAB “Proit”	Latvia
Umeå university	Sweden
Universidad de Alicante	Spain

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
Universidad de Jaén	Spain
Universidad de Murcia	Spain
Università Cattolica del Sacro Cuore	Italy
Università degli studi di Torino	Italy
Universität Hamburg	Germany
Universitat Jaume I	Spain
Universitat Politècnica de Catalunya	Spain
Universitat Pompeu Fabra	Spain
Université Parus-Saclay	France
University “Politehnica” of Bucharest	Romania
University of Alcalá	Spain
University of Amsterdam	Netherlands
University of Antwerp	Belgium
University of Belgrade	Serbia
University of Bergen	Norway
University of Brasília	Brazil
University of Bristol	UK
University of Coimbra	Portugal
University of Copenhagen	Denmark
University of Edinburgh	UK
University of Essex	UK
University of Gothenburg	Sweden
University of Groningen	Netherlands
University of Haifa	Israel
University of Helsinki	Finland
University of Jaén	Spain
University of Library Studies and Information Technologies	Bulgaria
University of Lisbon	Portugal
University of Ljubljana	Slovenia
University of Luxembourg	Luxembourg
University of Malta	Malta
University of Manchester	UK
University of Maribor	Slovenia
University of Nova Gorica	Slovenia
University of Patras	Greece
University of Pécs	Hungary
University of Porto	Lisbon
University of Primorska	Slovenia
University of Santiago de Compostela	Spain
University of Sheffield	UK
University of St-Etienne	France
University of Stuttgart	Germany
University of Szeged	Hungary
University of Tartu	Estonia
University of the Aegean	Greece
University of the Basque Country	Spain
University of Twente	Netherlands
University of Vienna	Austria
University of Vigo	Spain
University of Warsaw	Poland
University of West Bohemia	Czech Republic
University of Zagreb	Croatia
University of Zurich	Switzerland

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
University Politehnica Bucharest	Romania
University Ss. Cyril and Methodius	North Macedonia
Uppsala University	Sweden
Utrecht University	Netherlands
Vicomtech	Spain
Vilnius university	Lithuania
Visma	Norway
VÓCALI Sistemas Inteligentes	Spain
Vocapia Research	France
Vytautas Magnus University	Lithuania
Wikimedia Deutschland	Germany
WordFinder Software International AB	Sweden
Worldwide Bildungswerk	Germany
Wrocław University of Science and Technology	Poland
WWU Münster	Germany
Zurich University of Applied Sciences	Switzerland
KTH Royal Institute of Technology	Sweden
LT Users	
Acapela Group	Belgium
Accademia della Crusca	Italy
ADAPT Centre	Ireland
All Ukrainian National Culteral Moldovan Association	Ukraine
Association of Language Testers in Europe	UK
Amical Wikimedia	Spain
Aragonese Wikipedia	Spain
Archil Eliashvili Institute of Control Systems of Georgian Technical University	Georgia
ARTE G.E.I.E.	France
Atercin	Ireland
Athena Research Centre	Greece
ATI Technologies	Canada
Babeş-Bolyai University	Romania
Bangor University	UK
Basque Radio Television Public Group	Spain
Basque Wikipedia	Spain
BEIA	Austria
Bibliothèque universitaire des langues et civilisations	France
Bulgarian Academy of Sciences, Institute for Bulgarian Language	Bulgaria
Bulgarian Wikipedia	Bulgaria
Carpatho-Rusyns	
Catholic University Eichstätt-Ingolstadt	Germany
CBAC-WJEC	UK
CEE Spring	
Central State Office for the Development of the Digital Society	Croatia
Centre for Aromanian Language and Culture	Bulgaria
Centre for the Greek Language	Greece
Centro de Computação Gráfica	Portugal
CNRS	France
Council for the Maltese Language	Malta
Cornwall Council	UK
Croatian Academy of Sciences and Arts	Croatia
Croatian Association of Scientific and Technical Translators	Croatia

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
Croatian Parliament	Croatia
Cymdeithas Cyfieithwyr Cymru	UK
Danish Language Council	Denmark
Debagoieneko Mankomunitatea	Spain
Departament d'Educació	Spain
Directorate-General for Language Policy. Government of the Balearic Islands	Spain
Dublin City University	Ireland
ECSPM	Denmark
Edilic Association	France
Educational & Training Concepts	Greece
English Wiktionary	
Ensino Português no Estrangeiro	Portugal
Eurescom GmbH	Germany
Euroglossa d.o.o.	Croatia
European Culture and Technology Lab+, Technological University Dublin	Irelans
Euskal Irrati Telebista	Spain
Faculty of Science, University of Split	Croatia
Federal Lezghin National and Cultural Autonomy	Russia
Food Standards Agency	UK
Foras na Gaeilge	Ireland
FP CGIL, Spaciada sa bregùngia, RAS	Italy
Fran Ramovš Institute of the Slovenian Language	Slovenia
French Wikipedia	
German Research Center for Artificial Intelligence	Germany
Gimara Ltd	Finland
Global Link d.o.o.	Bosnia and Herzegovina
Globe	USA
Grow Coaching Alliance	Greece
Haute Ecole pédagogique Vaud	Switzerland
HGK	Germany
Hilfsgemeinschaft der Blinden und Sehschwachen Österreichs	Austria
Hitz Center (Ixa Research Group)	Spain
Hungarian Research Centre for Linguistics, Budapest University of Technology and Economics	Hungary
Hse	Germany
ICC-Languages	Germany
Institute for Language and Folklore	Sweden
Institute for Social Research in Zagreb	Croatia
Institute of Croatian Language and Linguistics	Croatia
Institute of Multilingualism at the University of Fribourg i.Ü.	Switzerland
Institute of the Estonian Language	Estonia
Institute of the Lithuanian Language	Lithuania
Instituto da Lingua Galega (Universidade de Santiago de Compostela)	Spain
Institutul de Filologie Română "A. Philippide" Academia Română , Romanian Academy	Romania
Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti", Academia Română , Romanian Academy	Romania
Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu", Academia Română , Romanian Academy	Romania
Joanneum Research	Austria

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
Intellectual Property Office	UK
International Medical Informatics Association (IMIA)	Switzerland
Interregional public association of Meskhetian Turks “Vatan”	Russia
Inuits	Greenland
IURIDICO Legal & Financial Translation sp. z o.o.	Poland
Lab University of Applied Sciences	Finland
Language Technologies Unit Bangor University	UK
Leibniz Institute for the German Language	Germany
LIBER	Netherlands
Library and Information Centre, Hungarian Academy of Sciences	Hungary
Librezale	Spain
LIDILE/Université Rennes 2	France
Lingua Libre	France
Linköping university	Sweden
Longbrook Translation	
LTU, Canolfan Bedwyr, Bangor University	UK
Macedonian Academy of Sciences and Arts	Macedonia
Macedonian Wikipedia	Macedonia
Maynooth university	Ireland
Media Perspectives	Netherlands
Megabyte Ltd	Malta
Menai Science Park Ltd	UK
Mercell	Norway
Ministère de l’éducation nationale	France
Ministry of Culture	
Ministry of Education	
Mirara Translations	Croatia
MITA	Malta
Nara Educational Technologies	USA
National and Kapodistrian University of Athens	Greece
National association of deaf women in Ireland	Ireland
National Research, Development and Innovation Office	
National Research Council of Italy	Italy
National Youth Service - Ministry of Education, Children and Youth	
NHS	UK
Nico van de Water Linguistic Services	Netherlands
Non profit ISSA Polska	Poland
NPLD	Belgium
Nuance Communication	USA
Open University of Catalonia	Spain
Pázmany Péter Catholic University	Hungary
Polytechnic University of Valencia	Spain
Projectus grupa	Croatia
Regione autonoma Valle d’Aosta	Italy
Research Center for Linguistics	Hungary
Research Centre of the Slovenian Academy of Sciences and Arts	Slovenia
RTL	Germany
Shell	UK
Smith& Nephew	UK
Spencer Stuart	USA
SPES d.o.o.	Slovenia
Språkrådet (The Language Council of Norway)	Norway
Staroslavenski institut (Old Church Slavonic Institute)	Croatia

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
Stockholm University	Sweden
storyfact.	
Tacawit Wiktionary	
Tampere University	Finland
Teaching council of Ireland	Ireland
Technical University of Denmark	Denmark
Telecats BV.	Netherlands
The Árni Magnússon Institute for Icelandic Studies	Iceland
The Finnish Social Insurance Institution	Finland
The Institute for the Languages of Finland	Finland
The Institute of the Lithuanian Language	Lituania
The National Library of Wales	UK
Top Communica	Poland
Toros University	Turkey
Trinity College Dublin	Ireland
Unesco, Digital Cultural heritage	
Universidad de Zaragoza	Spain
Universidade de Santiago de Compostela	Spain
Universitat Autònoma de Barcelona	Spain
Universitat Oberta de Catalunya	Spain
Universitat Politecnica de Valencia	Spain
Université Rennes 2	France
university college Dublin	Ireland
University of Athens	Greece
University of Bamberg	Germany
University of Belgrade, Faculty of Mining and Geology (Language Technology group)	Serbia
University of Bristol	UK
University of Cambridge	UK
University of Copenhagen	Denmark
University of Eastern Finland	Finland
University of Economics, Bratislava	Slovakia
University of Edinburgh	UK
University of Extremadura	Spain
University of Food Technologies - Plovdiv	Bulgaria
University of Győr	Hungary
University of Hamburg	Germany
University of Luxembourg	Luxembourg
university of Lyon	France
University of Malta	Malta
University of Osijek	Croatia
University of Padua	Italy
University of Porto	Portugal
University of Rijeka	Croatia
University of Strasbourg	France
University of The Basque Country	Spain
University of Thessaly	Greece
University of Vigo	Spain
University of York	UK
University of Zagreb, Faculty of Electrical Engineering and Computing	Croatia
University of Zagreb, Faculty of Humanities and Social Sciences	Croatia
Vilnius university	Lithuania

Continued on next page

Table 11 – *Continued from previous page*

Organisation	Country
Vytautas Magnus University	Lithuania
Washington Metropolitan University	USA
Wikidata	
Wikimedia Community Ireland	Ireland
Wikimedia Community User Group Malta	Malta
Wikimedia Denmark	Denmark
Wikimedia Deutschland e.V.	Germany
Wikimedia Foundation Search Platform Team	USA
Wikimédia France	France
Wikimedia Hungary	Hungary
Wikimedia UK	UK
Wikimedians of Slovakia	Slovakia
Wrocław University of Science and Technology	Poland
Y Coleg Cymraeg Cenedlaethol	UK
Zagreb School of Economics and Management	Croatia
Zurich University of Applied Sciences	Switzerland
Òmnium Cultural	Iceland