EUROPEAN LANGUAGE EQUALITY 2

D3.1

New strategic documents and technology in LT and language-centric AI

Authors	Itziar Aldabe, Aritz Farwell and German Rigau
Dissemination level	Public
Date	29-12-2022

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
Deliverable number	D3.1
Deliverable title	New strategic documents and technology in LT and language-centric AI
Type	Report
Number of pages	51
Status and version	Final
Dissemination level	Public
Date of delivery	29-12-2022
Work package	WP3: Strategic Research, Innovation & Deployment Agenda: Maintenance and Extension
Task	Task 3.1 Monitoring LT and language-centric AI
Authors	Itziar Aldabe, Aritz Farwell and German Rigau
Reviewers	Federico Gaspari, Georg Rehm
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2022 ELE2 Consortium

Consortium

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

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory Committee
CAGR	Compound Annual Growth Rate
CCS	Cultural and Creative Sectors
CEF	Connecting Europe Facility
CLARIN	Common Language Resources and Technology Infrastructure
COE	Council of Europe
CPAI	Coordinated Plan on Artificial Intelligence
CRACKER	Cracking the Language Barrier (EU project, 2015–2017)
CSA	Coordination and Support Action
DGA	Data Governance Act
DG CNECT	Directorate-General for Communications Networks, Content and Technology
DLE	Digital Language Equality
EC	European Commission
ECSPM	European Civil Society Platform for Multilingualism
ECRML	European Charter for Regional or Minority Languages
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality
ELE1	European Language Equality (preceding project)
ELE2	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELG	European Language Grid (EU project, 2019-2022)
ELIS	European Language Industry Survey
ELRC	European Language Resource Coordination
ELT	European Language Technology
EP	European Parliament
ESFRI	European Strategy Forum on Research Infrastructures

EU	European Union
GDPR	General Data Protection Regulation
HAI	Institute for Human-Centered AI (Stanford University)
HPC	High Performance Computing
IA	Innovation Action
IDB	Inter-American Development Bank
JRC	Joint Research Center
LT	Language Technology/Technologies
LLM	Large Language Models
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
MFF	Multiannual Financial Framework
ML	Machine Learning
MT	Machine Translation
NCC	National Competence Centre
NCP	National Contact Point
NLP	Natural Language Processing
NLU	Natural Language Understanding
OCLC	Online Computer Library Center
OECD	Organisation for Economic Co-operation and Development
PPP	Public-Private Partnership
R&D	Research and Development
R&D&I	Research and Development and Innovation
RIA	Research and Innovation Action
RI	Research Infrastructure
RML	Regional and Minority Languages
ROI	Return of Investment
SME	Small and Medium-sized Enterprises
SRA	Strategic Research Agenda
SRIA	Strategic Research and Innovation Agenda
STOA	Science and Technology Options Assessment

Abstract

As in ELE1, ELE2 continues the process of monitoring new international, national and regional Strategic Research Agendas (SRAs), studies and initiatives related to Language Technology (LT) and Artificial Intelligence (AI) as well as recent breakthroughs in LT and language-centric AI technology as a whole. ELE2 continues to revise key research areas and gaps in research that need to be addressed to ensure that the current serious inequality in LT support for all Europe's languages can be overcome.

This document reports on the continuous desk research towards the systematic collection and analysis of the existing international, national and regional SRAs, studies, reports and initiatives related to LT and LT-related AI. This document also reports on the latest breakthroughs in the field of LT and language-centric AI. This task will continue until the end of the project in June 2023 in order to extend the current ELE1 SRIA to ensure that Digital Language Equality (DLE) in the EU becomes a reality by 2030.

Around 200 documents from within the European Union (EU) and international sources were reviewed and analysed in ELE1 for this purpose. The original ELE1 Deliverable 3.1 *Report on existing strategic documents and projects in LT/AI*, released in April 2021, was updated twice, first in November 2021 and then again in April 2022 (Aldabe et al., 2022).[1]

This document also reports on the most recent breakthroughs in the field of LT and language-centric AI since the ELE1 Deliverable 1.2 *Report on the state of art in Language Technology and Language-centric AI* released in September 2021 reported on the current state of the art in the field of LT and language-centric AI. The main purpose of this deliverable was to landscape the field of LT and language-centric AI by assembling a comprehensive report on the state of the art of basic and applied research in the area (Agerri et al., 2021).[2]

The collected AI and LT documents, reports and initiatives are listed in Appendix A.[3]

1 Introduction

1.1 Languages in Europe

In varietate concordia (*united in diversity*) 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"[4] (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,[5] 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[6] states that "[a]ny 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).

[1] https://european-language-equality.eu/wp-content/uploads/2022/06/ELE___Deliverable_D3_1__second_revision_2.pdf

[2] https://european-language-equality.eu/wp-content/uploads/2022/06/ELE___Deliverable_D3_1__second_revision_2.pdf

[3] Links and URLs mentioned in the report last accessed online as of December of 2022.

[4] http://europa.eu/abc/symbols/motto/index_en.htm

[5] https://en.m.wikipedia.org/wiki/European_Charter_for_Regional_or_Minority_Languages

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

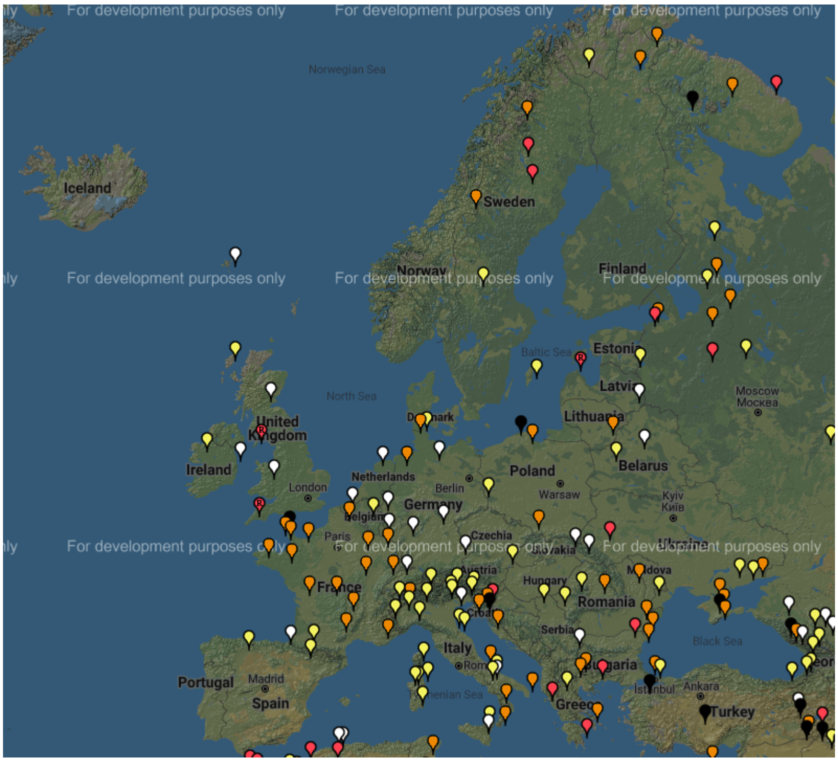

Figure 1: Endangered European languages according to the UNESCO Atlas of the World's Languages in Danger.

As is clear, multilingualism is one of the key cultural cornerstones of Europe and signifies a large part of what it means to be and to feel European. However, not only do language barriers still hamper cross-lingual communication and the free flow of knowledge and thought across borders, many languages themselves are endangered or on the edge of extinction. Figure 2 shows the languages in danger in Europe according to the *UNESCO Atlas of the World's Languages in Danger* (Moseley, 2010).[7] In this map, black flags correspond to already extinct languages. A new version of this Atlas is expected to be released soon.[8]

Furthermore, no common EU policy has been proposed to address the problem of language barriers. The EU's long-term budget for 2021-2027, coupled with NextGenerationEU, the temporary instrument designed to boost the recovery after the COVID-19 pandemic, 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.[9] 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).[10]

[7] https://unesdoc.unesco.org/ark:/48223/pf0000187026

[8] http://www.unesco.org/languages-atlas/

[9] https://ec.europa.eu/info/strategy/recovery-plan-europe_en

[10] https://digital-strategy.ec.europa.eu/en/activities/digital-programme

1.2 Language-centric Artificial Intelligence

Because natural language is at the heart of human intelligence, it is and must be at the heart of our efforts to develop AI technologies.[11] No sophisticated and effective AI-powered tool can exist without mastery of language.[12] Thus, language is the next great frontier in AI.[13] In fact, LT is already arguably the hottest field of AI.[14] Together with vision and robotics, several recent international reports place LT as one of the three core application areas within AI. Automatic language understanding is perceived as one of the fundamental goals of AI and, in turn, is also considered one of its main challenges (Sayers et al., 2021). Over the years, LT has developed different methods to make the information contained in written and spoken language explicit or to generate written and spoken language. Despite the inherent difficulty of many of the tasks performed, current LT support allows for many advanced applications which would have been unthinkable only a few years ago. Today, many people use LT on a regular basis, especially online, mostly without even knowing that they do. LT is an important but often invisible ingredient of applications as diverse as search engines, spellcheckers, text editors, text predictors, machine translation (MT) systems, recommender systems, virtual assistants, chatbots, automatic subtitling, automatic summaries, inclusive technology, voice synthesizers and many others. It is the nerve centre 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. Its rapid development promises even more encouraging and exciting results in the near future.

LT is multidisciplinary since it combines knowledge in computer science (and specifically in AI), mathematics, linguistics and psychology among others. Figure 2 shows some of the most important disciplines involved in LT. This uniqueness must be considered in any public or private initiative devoted to AI that includes LT. Only the proper application of LT will allow processing and understanding, i. e., making sense of these enormous volumes of multilingual written, spoken and visual data in sectors as diverse as health, justice, education, or finance. LT applications such as speech recognition, speech synthesis, textual analysis and MT are used by hundreds of millions of users on a daily basis.[15] As reflected in national and regional AI and LT strategies both inside and outside Europe, as well as in its inclusion in all the prioritized strategic areas for developing research, development and innovation (R&D&I) activities, LT is highlighted as one of the most relevant technologies for society.

1.3 Content and structure of this document

This document provides an in-depth review and analysis of the existing international, national and regional SRAs, studies, reports and initiatives related to LT and AI. Close to 200 such documents from within the EU and international sources have been reviewed and analysed for this purpose.[16]

The structure of this document is as follows. Section 2 provides a brief historical overview of the LT area. Section A reviews the reports produced by the main international organizations regarding AI and LT. Section 4 focuses on the main European SRAs, initiatives and National Plans regarding AI and LT. In order to position language-centric AI in Europe, Section 5 analyses its strengths and weaknesses, opportunities and threats, showing its unique and multidisciplinary nature in terms of the issues addressed. Finally, Section 6 summarizes a series of recommendations for strengthening LT in Europe.

[11] https://hbr.org/2022/04/the-power-of-natural-language-processing

[12] https://www.nytimes.com/2022/04/15/magazine/ai-language.html

[13] https://www.forbes.com/sites/robtoews/2022/02/13/language-is-the-next-great-frontier-in-ai/?sh=6995a0865c50

[14] https://analyticsindiamag.com/is-nlp-innovating-faster-than-other-domains-of-ai/

[15] https://www.nimdzi.com/nimdzi-language-technology-atlas-2020/

[16] All collected AI and LT documents, reports and initiatives are listed in Appendix A

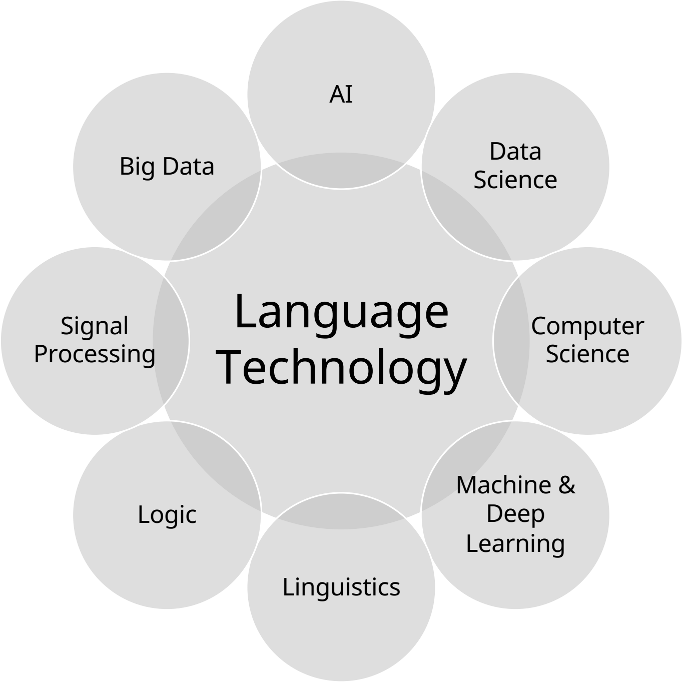

Figure 2: Language Technology as a multidisciplinary field.

2 Language Technology: General Overview

2.1 A very brief historical view

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 high-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 what it perceived to be more realistic objectives, especially in the short term.

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) machines 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., 2013; 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 Graphics Processing Units, 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.

2.2 The Deep Learning era

In this present era, LT has been moving away from a methodology in which a pipeline of multiple modules is utilized to implement LT solutions to architectures based on complex neural networks trained with vast amounts of text data. Moreover, as highlighted by the 2021 *AI Index Report*,[17] deep learning techniques have simultaneously engendered rapid progress in NLP, vision and robotics. The reasons behind the advances are gathered in a separate report issued by the Joint Research Centre, *Artificial Intelligence: A European Perspective*,[18] which underscores that success in these areas of AI has been possible because four research trends have converged: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual textual 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; Min et al., 2021a). Put together, these currents have helped bring about a new state of the art through systems that apparently obtain human-level performance in laboratory benchmarks when testing difficult English-language understanding tasks.

In some cases, an increase in scale has led to such behaviour. 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).[19] Not only was this ability mostly absent from its predecessor GPT-2, over 100 times smaller than GPT-3, the latter also outperforms state-of-the-art models on certain tasks for which they were explicitly trained to solve. Furthermore, recent work has demonstrated that pretrained language models can robustly perform NLP 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). Surprisingly, fine-tuning pretrained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Aghajanyan et al., 2021; Aribandi et al., 2021; Min et al., 2021b; Sanh et al., 2021; Wei et al., 2021; Ye et al., 2021). 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 have encouraged various IT enterprises, including Google, Microsoft and OpenAI, to develop and deploy their own large pretrained neural language models.

[17] https://aiindex.stanford.edu/report/

[18] https://publications.jrc.ec.europa.eu/repository/handle/JRC113826

[19] 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/).

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.[20] To begin with, we currently have no clear understanding of how they work, when they fail, or 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 large and very well-resourced organisations can currently afford to construct them. 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, or source data that deals with a number of other minorities (Dodge et al., 2021). In addition, the restricted access mentioned above limits researchers' ability to understand how and why large language models work, hindering progress to improve their robustness and mitigate known issues of toxicity.[21] Moreover, large language models can sometimes produce unpredictable and factually inaccurate text or even recreate private information.[22] Fifth, computing large pretrained models comes with a substantial carbon footprint. Strubell et al. (2019) 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.[23] Finally, the implications of *foundation models* may extend to questions of AI nationalism as well. Several countries appear to be engaging in an "AI arms race", prompting the MIT Technology Review to call 2021 "the year of monster AI models", which include language and multimodal models that have come out of the US, China, South Korea, and Israel.[24] In short, notwithstanding claims of human parity in many LT tasks, Natural Language Understanding (NLU) is still an *open research problem* far from being solved since all current approaches have *severe* limitations.

Given these issues and the role of LT in everyone's daily lives, many LT practitioners are particularly concerned by one area within the field: language diversity in LT research and the need for transparent DLE across all aspects of European society, from government to business to citizen.[25] 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; Blasi et al., 2022). Joshi et al. (2020), for instance, examine the relationship between types of languages, resources and their representation in NLP conferences over time. As expected, only a small number of the world's 7000+ languages are represented in the rapidly evolving LT field. This disproportionate representation is further exacerbated by systematic inequalities in LT across the world's lan-

[20] https://lastweekin.ai/p/the-inherent-limitations-of-gpt-3

[21] https://www.unite.ai/minority-voices-filtered-out-of-google-natural-language-processing-models/; https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/

[22] https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html

[23] https://ourworldindata.org/co2-emissions

[24] https://www.technologyreview.com/2021/12/21/1042835/2021-was-the-year-of-monster-ai-models/; https://lastweekin.ai/p/gpt-3-foundation-models-and-ai-nationalism?s=r

[25] https://gitlab.com/ceramisch/eacl21diversity/-/wikis/EACL-2021-language-diversity-panel

guages. 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's development is driven by the economic status of a language's users, rather than 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 a path to leverage LT for less-developed languages. 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 will require resources (experts, data, computing facilities, etc.) along with carefully designed evaluation benchmarks and annotated datasets for every language and domain of application.

What other measures must be taken to ensure LT's multifaceted potential is fully realized? 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),[26] produce playscripts (Rosa et al., 2020), yield text that is nearly indistinguishable from human prose (Brown et al., 2020), and successfully solve unseen tasks (Wei et al., 2021; Sanh et al., 2021; Min et al., 2021b; Ye et al., 2021; Aghajanyan et al., 2021; Aribandi et al., 2021). 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. Hence, as this brief history and overview of LT make clear, the moment to assess its present status and chart its future course is now.

2.3 Economic impact of LT

Unsurprisingly, LT is also one of the fastest growing application areas in AI. Funding for LT start-ups is booming.[27] Early-stage funding in 2021 amounted 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.[28] 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,[29] the global NLP market size is expected to grow from USD 10.2 billion in 2019 to USD 26.4 billion by 2024, at a CAGR (Compound Annual Growth Rate) of 21% is set during the forecast period 2019-2024.[30] According to another

[26] https://openai.com/blog/dall-e/

[27] https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming

[28] https://towardsdatascience.com/nlp-how-to-spend-a-billion-dollars-e0dcdf82ea9f

[29] https://www.businesswire.com/news/home/20191230005197/en/Global-Natural-Language-Processing-NLP-Market-Size

[30] https://www.analyticsinsight.net/potentials-of-nlp-techniques-industry-implementation-and-global-market-outline/

report from the end of 2019,[31] 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% 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 by 2026, exhibiting a CAGR of 32.4% during the forecast period.[32] Another report from 2020 estimates the global LT market to reach USD 41 billion by 2025.[33] In a recent 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.[34] Another report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at 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.[35] A subsequent report published in 2021 estimated that the global NLP market is predicted to grow from USD 20.98 billion in 2021 to USD 127.26 billion in 2028 at a CAGR of 29.4% in the forecasted period.[36] NLP in Europe will witness market growth of 19.7% CAGR and is expected to reach USD 35.1 billion by 2026.[37] As a final example, according to Straits Research the global NLP market size was worth USD 13.5 billion in 2021 and it is estimated to reach an expected value of USD 91 billion by 2030, growing at a CAGR of 27% during the forecast period (2022–2030).[38] These numbers indicate that the return of investment (ROI) will be massive.

3 Language Technology in International Organizations

AI capabilities are rapidly evolving, shaping the quick evolution of one of the 21st century's most transformative technologies.[39] The growing interest in AI at a global political, scientific and social level has led several international organizations to draft a number of reports and initiatives in recent years. These often focus on the socioeconomic impact of AI technologies and applications with respect to policy.

3.1 Reports from International Organizations

The Organisation for Economic Co-operation and Development (OECD),[40] a frequent contributor to this discourse, has helped coordinate dialogue on the subject at international fora (notably the G7, G20, EU and UN), offered practical advice to governments on how to actualize AI policy, and stressed the potential that digital technologies demonstrate in responding to societal challenges.[41] Its 2021 report, *State of the implementation of the OECD AI Principles: Insights from national AI policies*, identifies challenges and best practices for the implementation of the five policy recommendations to national governments contained in

[31] https://www.vynzresearch.com/ict-media/natural-language-processing-nlp-market

[32] https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933

[33] 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

[34] 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

[35] https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market

[36] https://www.analyticsinsight.net/the-global-nlp-market-is-predicted-to-reach-us127-26-billion-by-2028/

[37] https://www.analyticsinsight.net/nlp-in-europe-is-expected-to-reach-us35-1-billion-by-2026/

[38] https://www.globenewswire.com/en/news-release/2022/08/11/2497065/0/en/Natural-Language-Processing-Market-Size-is-projected-to-reach-USD-91-Billion-by-2030-growing-at-a-CAGR-of-27-Straits-Research.html

[39] https://www.holoniq.com/notes/50-national-ai-strategies-the-2020-ai-strategy-landscape/

[40] https://www.oecd.org

[41] See, e. g., *Artificial Intelligence in Society* (https://doi.org/10.1787/eedfee77-en); *State of the implementation of the OECD AI Principles: Insights from national AI policies* (https://doi.org/10.1787/1cd40c44-e); *The Digitalisation of Science, Technology and Innovation* (https://doi.org/10.1787/b9e4a2c0-en).

its OECD AI Principles. These are: 1) invest in AI R&D; 2) foster a digital ecosystem for AI; 3) shape an enabling policy environment for AI; 4) build human capacity and preparation for labour market transformation; and 5) foment international co-operation for trustworthy AI. The report comes on the heels of the OECD's *The Digitalisation of Science, Technology and Innovation*, which emphasises that cutting-edge NLP techniques are opening new analytical possibilities. Among those listed is the ability to recognize victims of sexual exploitation on the internet based on facial detection and social network analysis (Chui et al., 2018). Advances such as this have caught the attention of researchers and policy makers in various countries, who have begun to experiment with NLP to track emerging research topics and technologies. As the report underscores, policy makers use these results to formulate science and innovation policy initiatives, support investments in R&D&I, and evaluate public programmes.[42]

Similar policy guidance and assessments appear elsewhere as well.[43] The World Economic Forum,[44] which provides a framework for governments that wish to develop national AI strategies, assists those responsible for crafting policy in how to ask pertinent questions, follow best practices, identify and involve stakeholders, and create a set of outcome indicators.[45] UNESCO[46] extends these considerations to the educational sphere, recommending that governments and other stakeholders, in accordance with their legislation and public policies, respond to education-related opportunities and challenges presented by AI. The *Beijing Consensus on Artificial Intelligence and Education*, an outcome document issued by UNESCO in 2019, stresses the multidisciplinary nature of AI and urges readers to consider the role of AI tools in teaching and learning, highlighting its effectiveness in aiding students with learning impairments or who study in a language other than their mother tongue.[47] In the area of library science, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*, a research position paper from OCLC,[48] notes several structural inequalities are perpetuated by data-driven policies (Padilla, 2020) and sets an agenda for tackling both positive and negative impacts of data science, machine learning, and AI on libraries.[49]

Finally, in March 2022, based on the report *Facilitating the implementation of the European Charter for Regional or Minority Languages through artificial intelligence* first published in February 2020[50] and updated in March 2022[51], the Committee of Experts of the European Charter for Regional or Minority Languages of the Council of Europe (CoE) adopted a statement on the promotion of regional or minority languages through artificial intelligence (AI).[52]. 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.

[42] To help policy makers, regulators, legislators and others characterise AI systems deployed in specific contexts, the OECD has developed a user-friendly tool to evaluate AI and LT systems from a policy perspective (https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm).

[43] See, e. g., the *Helsinki Initiative on Multilingualism in Scholarly Communication* (https://www.helsinki-initiative.org/en).

[44] https://www.weforum.org

[45] https://www3.weforum.org/docs/WEF_National_AI_Strategy.pdf

[46] https://en.unesco.org

[47] https://unesdoc.unesco.org/ark:/48223/pf0000368303 See also, UNESCO's *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*, a 2019 report which, among other breakthroughs, noted a Chinese AI system that is able to correct student essays as a milestone in LT for education (https://unesdoc.unesco.org/ark:/48223/pf0000366994).

[48] https://www.oclc.org/en/about.html

[49] https://doi.org/10.25333/xk7z-9g97

[50] https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

[51] https://rm.coe.int/min-lang-2022-4-ai-and-ecrml-en/1680a657c5

[52] https://rm.coe.int/declaration-ai-en/1680a657ff

The Inter-American Development Bank[53] (IDB) also advocates for building a shared understanding of AI, its opportunities and applications, as well as its risks and the possible measures to mitigate them.[54] The referenced subjects include AI, computer vision, machine learning and data mining, and NLP.

The attention paid to AI and LT in policy reports reflects the social, political, and economic importance that the technology has garnered in today's world; and the same holds true for organizations that trace trends in innovation. In its report, *Technology Trends 2019 Artificial Intelligence*,[55] the World Intellectual Property Organization[56] found that 50% of all AI patents have been published in just the past five years, a striking illustration of how rapidly innovation is advancing in the field. The report, which classifies AI technology trends into techniques, functional applications, and application fields, furthermore points to LT as one of AI's most significant functional applications, attributing over a quarter of all AI-related patents to NLP and speech processing. The number is unsurprising given NLP's current incandescence within AI, where the rising star is turning many heads. A case in point is the *State of AI Report* for 2021,[57] issued by UK AI investors with an eye toward stimulating informed conversation on AI and its implications going forward. The report, which considers research, talent, industry, and politics, discusses the emergence of large language models and notes that the latest generation are unlocking new NLP use cases. Indeed, the arrival of Transformers as a general purpose architecture for ML has been a revelation, beating the state of the art in domains as disparate as computer vision and protein structure prediction.

3.2 Reports from the United States

Reports from the United States tell an analogous story to their international counterparts. In its 2021 and 2022 *AI Index Reports* (Zhang et al., 2021), for instance, the Institute for Human-Centered AI (HAI) at Stanford University reviews the growth of research papers and conferences over time and by region, tracks AI accuracy on several benchmarks, focuses on trends in jobs and investment, and examines various national AI strategies.[58] The reports also devote space to data and analysis concerning AI with respect to education, diversity, and ethics. Key takeaways include the observation that 65% of the new PhDs in the US chose jobs in industry over academia compared to 44.4% the previous year, that there is still little data available on the ethical challenges surrounding AI, and that the AI workforce remains predominantly male. The 2022 report also highlights that while current large language models are setting records on technical benchmarks, they are also increasingly reflecting biases from their training data (multimodal models learn multimodal biases, for instance). The reports' findings are accompanied by HAI's updated Global Vibrancy Tool,[59] which measures performance on various economic, inclusion, and R&D factors across several countries. The tool can create an overall index for the full list of 26 countries and it is of note that none of the top ten is an EU member state. The worrisome nature of the latter data point is compounded in an examination of the global balance and flow of top AI scientists provided by the Paulson Institute's Macro Polo think tank in its Global AI Talent Tracker report.[60] According to this analysis, the US lead in AI is built on attracting international talent, with more than

[53] https://www.iadb.org/en

[54] https://publications.iadb.org/en/artificial-intelligence-for-social-good-in-latin-america-and-the-caribbean-the-regional-landscape-and-12-country-snapshots

[55] https://www.wipo.int/publications/en/details.jsp?id=4386

[56] https://www.wipo.int

[57] https://www.stateof.ai

[58] The 2021 report is divided into seven chapters: Research and Development; Technical Performance; The Economy; AI Education; Ethical Challenges of AI Applications; Diversity in AI; and AI Policy and National Strategies (https://aiindex.stanford.edu/report/).

[59] https://aiindex.stanford.edu/vibrancy/

[60] https://macropolo.org/digital-projects/the-global-ai-talent-tracker/

two-thirds of the top-tier AI researchers working in the US having received undergraduate degrees in other countries. Although 18% of the top-tier AI researchers are European, only 10% of them work in Europe.

These final details should sound alarm bells in Europe. As demonstrated in *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence*, released by the AI100 project in 2021, remarkable progress has been made in AI over the past five years and we may anticipate that its effects will ripple out for many years to come. Prepared by a panel of experts from around the globe, the report makes clear that the ability of computers to perform sophisticated language- and image-processing tasks has advanced significantly and that more investment of time and resources is required to meet the challenges posed by AI's rapidly evolving technologies. On the one hand, this includes greater government involvement in the areas of regulation and digital education. In an AI-enabled world, citizens young and old must be literate in these new digital technologies. On the other, this means addressing fears that AI technologies will contribute to unemployment in some sectors. A Blumberg Capital survey of 1,000 American adults found that about half are concerned that AI threatens their livelihood. Indeed, despite the fact that 72% agreed that AI would help remove tedious tasks and free up time to concentrate on more creative ones, 81% were reluctant to surrender these tasks to an algorithm for fear of being supplanted.[61] As the authors of Gathering Strength, Gathering Storms indicate, AI is leaving the laboratory and entering our lives, having a "real-world impact on people, institutions, and culture."[62]

This perspective is shared by the National Security Commission. In addition to raising concerns that the United States risks falling behind China and other countries in the AI race, its recent 750-page report encourages the federal government to step up investment in the area.[63] Specifically, the commission calls for a *modest down payment* of $40 billion, along with hundreds of billions more in the coming years, to galvanize future breakthroughs and help democratize AI research. Moreover, the report provides policy makers with a guide to ensure the US is prepared to defend against AI threats, promote AI innovation, and make responsible use of AI for national security. It is also worth mentioning that the report lists NLU as one of the six uses for deployed AI today. This view, which coincides with the general consensus on LT expressed above, is further reinforced by the Future Today Institute[64] in its 2021 *Tech Trends Report* on AI.[65] The group not only identifies NLP as an area that is experiencing high interest, investment, and growth, but also forecasts that NLP algorithms will do more in the future, including, for example, aid in interpreting genetic changes in viruses.

3.3 Reports from the European Union

Reports from the European Union paint an equally upbeat picture about present and future expectations regarding science and technology. A recently released Eurobarometer survey on European citizens' knowledge and attitudes towards these shows that 86% believe the overall influence of science and technology is positive.[66] EU citizens expect a range of technologies currently under development, including AI (61%), to improve their way of life over the next 20 years. The case for AI and LT is further laid out by various European institutions in several recently issued reports and policy initiatives that highlight their extensive impact on society and what must be done to shepherd this influence. For instance, *European Artifi-*

[61] https://blumbergcapital.com/ai-in-2019/
[62] https://ai100.stanford.edu
[63] https://www.nscai.gov/2021-final-report
[64] https://futuretodayinstitute.com
[65] https://2021techtrends.com/AI-Trends
[66] https://europa.eu/eurobarometer/surveys/detail/2237

cial Intelligence (AI) leadership, the path for an integrated vision[67]; *the Strategy on AI*[68]; *Ethics Guidelines for Trustworthy AI*[69]; *Liability for AI and other emerging technologies*[70]; *On Artificial Intelligence: A European approach to excellence and trust*[71]; *Coordinated Plan on AI*[72]. All agree that AI is an area of strategic importance, a key driver of economic development, and a means to provide solutions to many societal challenges. As such, they concur that the socioeconomic, legal and ethical impact of AI must be carefully weighed. For instance, the Joint Research Center (JRC) Science for Policy report, *The Changing Nature of Work and Skills in the Digital Age*,[73] observes that employment opportunities related to the development and maintenance of AI technologies and Big Data infrastructures are expected to grow, whereas jobs that are most vulnerable to automation appear to be those that require relatively low levels of formal education, do not involve complex social interaction, or demand routine manual tasks. Keeping this range of possibilities in mind is a reminder that digital technologies may not only create or destroy some lines of work, but also fundamentally change what people do on the job and how they do it.

The European Commission's new Coordinated Plan on AI states that that NLP is one of the most rapidly advancing fields within AI, and is designed in part to address such potential turbulence.[74] The 2021 plan, in conjunction with the first-ever legal framework for AI,[75] will guarantee the safety and essential rights of people and businesses, while strengthening AI uptake, investment and innovation across the EU. It is also seen as the EU's next step in fostering global leadership in trustworthy AI, deemed necessary if European AI is to be globally competitive while respecting European values. This is of particular concern given that the EC's 2021 Strategic Foresight Report, *The EU's capacity and freedom to act*,[76] stresses the EU's capabilities in AI, Big Data and robotics lag behind the world's leaders, the US and China. To strengthen European AI and digital sovereignty, the report encourages stakeholders to promote values via the finance, development and production of next-generation tech. One important area of focus must be high-value data, a key factor in improving performance and building robust AI models. The EC wants to ensure legal clarity in AI-based applications, especially regarding data. Its proposed regulation on data governance will help by boosting data sharing across sectors and member states, while the General Data Protection Regulation (GDPR) is a major step towards building trust.[77] The member states also agreed to a negotiating mandate on a proposal for a Data Governance Act (DGA),[78] part of a wider policy to give the EU a competitive edge in the increasingly data-driven economy. The aim is to promote the availability of data that can be utilized to power applications and advanced solutions in AI, personalised medicine, green mobility, smart manufacturing and numerous other areas. While these regulations support the privacy and rights of European citizens, it should be pointed out that significant barriers to the access and re-use of language resources remain, especially with regard to competition with countries that adopted the “fair use” doctrine, such as the US, Japan or Korea.

The EU's approach to data must be crafted with Big Data technology and LT in mind. The

[67] https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2018)626074

[68] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence#Building-Trust-in-Human-Centric-Artificial-Intelligence

[69] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

[70] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199

[71] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[72] https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence

[73] https://publications.jrc.ec.europa.eu/repository/handle/JRC117505

[74] https://digital-strategy.ec.europa.eu/en/library/new-coordinated-plan-artificial-intelligence

[75] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

[76] https://ec.europa.eu/info/strategy/strategic-planning/strategic-foresight/2021-strategic-foresight-report_en

[77] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[78] https://www.consilium.europa.eu/en/press/press-releases/2021/10/01/eu-looks-to-make-data-sharing-easier-council-agrees-position-on-data-governance-act/

latest roadmap from the European Strategy Forum on Research Infrastructures (ESFRI) includes Big Data technology as one of the emerging drivers of landscape analysis. According to its findings, research infrastructures in LT are indispensable in breaking new ground because they represent a core aspect of Big Data technology due to the volume and variety of data generated by the accumulation of unstructured text. Among the relevant research infrastructures is the Common Language Resources and Technology Infrastructure (CLARIN), an ESFRI Landmark and ERIC which offers interoperable access to language resources and technologies for researchers in the humanities and social sciences through a single online environment. Unfortunately, not every EU Member State is officially affiliated with it, while others participate only as observers (Belgium joined CLARIN in September 2021 and Spain will join in 2023). Additionally, because research funding agencies provide unbalanced resources to the different member states, European languages are not equally supported by CLARIN (de Jong et al., 2020). This aspect has received more attention in the EU project European Language Grid (ELG), which started in January 2019 and concluded in June 2022. The ELG cloud platform contains more than 14,000 language services and language resources for all European languages (Rehm et al., 2021; Rehm, 2023).[79] And as the main task in AI's communication domain, NLP encompasses applications such as text generation, text mining, text classification, MT and speech recognition. Put differently, LT's ability to analyze, understand and generate information expressed in natural language is crucial for improving human-computer interaction. This view is confirmed by AI Watch, the EC's knowledge service responsible for monitoring the development, uptake and impact of AI, in three recent reports, *Defining Artificial Intelligence*, *Artificial Intelligence in public services* and *AI Watch, road to the adoption of Artificial Intelligence by the public sector*.[80] By way of example, the latter identified and employed 230 cases of AI usage in public services in order to extract emerging trends in AI, revealing that well over half of the cases are closely related to LT.

Relatedly, the EC's Directorate-General for Communications Networks, Content and Technology (DG CNECT), in collaboration with the Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (DG GROW), opened a consultation in 2021 that examined use cases for website translation at small and medium-sized enterprises (SMEs) and surveyed multilingual websites in an effort to analyse language barriers across EU Member States.[81] The inquiry identified specific market needs that could be addressed through public solutions, such as eTranslation, and by European language service providers. Of the over 1,000 SMEs that responded, 75% expressed interest in participating in the EC's subsequent pilot program to make their website automatically multilingual. When the *European Language Industry Survey* (ELIS)[82] – then known as the *EUATC survey* – was run for the first time in 2013, MT was still primarily seen as a threat and a challenge; only a few language companies saw it as an opportunity. Today 65% of language company respondents see the improved quality of neural MT as an opportunity rather than a threat. According to the 2022 survey, 58% of those companies have implemented the technology and an additional 20% are planning to do so. This potential willingness to incorporate LT and AI corresponds with a separate study conducted by Eurostat[83] in 2020, which found that 7% of EU enterprises with at least ten employees used AI applications; 2% utilized ML to analyse big data internally and 1% evaluated big data internally with the help of LT. Moreover, 2% provided a chat service, where a chatbot or virtual agent generated natural language replies to customers.

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

[80] https://knowledge4policy.ec.europa.eu/ai-watch_en; https://publications.jrc.ec.europa.eu/repository/handle/JRC118163; https://publications.jrc.ec.europa.eu/repository/handle/JRC120399; https://joinup.ec.europa.eu/collection/innovative-public-services/news/ai-watch-road-adoption-artificial-intelligence

[81] https://digital-strategy.ec.europa.eu/en/library/report-sme-survey-multilingual-websites

[82] https://elis-survey.org

[83] https://ec.europa.eu/eurostat/web/main/home

4 Language Technology in Europe

4.1 European initiatives

The European Parliament officially took a very clear stance emphasizing that "multilingualism presents one of the greatest assets of cultural diversity in Europe and, at the same time, [is] one of the most significant challenges for the creation of a truly integrated EU."[84] The belief is reflected in the EU's promotion of multilingualism, which falls within the scope of a variety of EU policy areas. While many of the multifaceted efforts to support Europe's languages are bearing fruit, still greater attention must be paid to removing barriers to intercultural and inter-linguistic dialogue as a means to stimulate mutual understanding. One means to achieve this is through language technology. However, although official EU languages are granted equal status politically, they are far from equally supported technologically.[85]

Several key strategy documents have contributed to the European debate on this subject in the past decade, including *The FLaReNet Strategic Language Resource Agenda* (Soria et al., 2014), *META-NET Strategic Research Agenda for Multilingual Europe 2020* (Rehm and Uszkoreit, 2013; Rehm et al., 2014a, 2016), *Language Technologies for Multilingual Europe: Towards a Human Language Project* (Rehm, 2017), and the STOA report, "Language Equality in the digital age: Towards a Human Language Project" (STOA, 2017). The latter helped pave the way for the preparation of the European Parliament's joint ITRE/CULT resolution, *Language equality in the digital age* (European Parliament, 2018),[86] adopted in a plenary meeting on 11 September 2018 with an overwhelming majority of 592 votes in favour, 45 against and 44 abstentions.

Approval of the resolution by such a wide margin demonstrates the importance and relevance of the issue. It includes more than 40 recommendations, structured into the four sections: "Improving the institutional framework for language technology policies at EU level", "Recommendations for EU research policies", "Education policies to improve the future of language technologies in Europe" and "Language technologies: benefits for both private companies and public bodies". Among the most salient items may be found the following (all emphases added; some items partially abbreviated):

- "recommends that in order to raise the profile of language technologies in Europe, the Commission **should allocate the area of 'multilingualism and language technology' to the portfolio of a Commissioner**; considers that the Commissioner responsible **should be tasked with promoting linguistic diversity and equality at EU level**, given the importance of linguistic diversity for the future of Europe;" (item 14)
- "suggests **ensuring comprehensive EU-level legal protection for the 60 regional and minority languages**, recognition of the collective rights of national and linguistic minorities in the digital world, and mother-tongue teaching for speakers of official and non-official languages of the EU;" (item 15)
- "calls on **the Member States to develop comprehensive language-related policies and to allocate resources and use appropriate tools in order to promote and facilitate linguistic diversity and multilingualism in the digital sphere**; stresses the **shared responsibility of the EU and the Member States** and in developing databases and translation technologies for all EU languages, including languages that are less

[84] *Language equality in the digital age*, European Parliament, 2018.

[85] See, e. g., *META-NET White Paper Series: Europe's Languages in the Digital Age* (Rehm and Uszkoreit, 2012; Rehm et al., 2014b); Rehm et al. (2020b); Rehm and Hegele (2018).

[86] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html

widely spoken; calls for **coordination between research and industry** with a common objective of enhancing the digital possibilities for language translation and with open access to the data required for technological advancement;" (item 17)

- "calls on the Commission **to establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels**, tailored specifically to Europe's needs and demands; emphasises that the programme should seek to tackle **deep natural language understanding** and 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 to reduce the technology gap between European languages; stresses that this should be done with the participation of research centres, academic, enterprises […] and other relevant stakeholders;" (item 25)
- "believes that […], **European education policies should be aimed at retaining talent in Europe**, should analyse the current educational needs related to language technology […]] and, based on this, **provide guidelines for the implementation of cohesive joint action at European level**, […], including the language-centric artificial intelligence industry; (item 34)
- "points to the need **to promote the ever-greater participation of women in the field of European studies on language technologies**, as a decisive factor in the development of research and innovation;" (item 36)

To these recommendations the remarks made by EC Commissioner Corina Crețu in her closing statement at the hearing on the resolution may be added:

> Ensuring appropriate technological support for all European languages will […] create jobs, growth and opportunities in the DSM. It will enhance the quality of public services, and reinforce a stronger sense of unity and belonging throughout Europe. […] under the next Multiannual Financial Framework (MFF), we will need to reinforce funding, research and education actions. […] overcoming language barriers in the digital environment is essential for an inclusive society, a vibrant DSM and for unity in diversity.

Commissioner Crețu's statement is in line with previous public appeals voiced in 2016 by former European Commission Vice President Andrus Ansip and in 2017 by Director General Roberto Viola (DG Connect) for the need to strengthen multilingualism through technologies.[87]

Following the two previous STOA events on language technologies in the EU (the first in 2013[88] and the second in 2017[89]), on 8th November 2022, the European Parliament's Panel for the Future of Science and Technology (STOA) hosted a workshop entitled *Towards full digital language equality in a multilingual European Union*.[90] The event brought together policy-makers and experts from academia and industry to discuss the current challenges

[87] See *How multilingual is Europe's Digital Single Market?* (https://ec.europa.eu/commission/commissioners/2014-2019/ansip/blog/how-multilingual-europes-digital-single-market_en); *Multilingualism in the Digital Age: a barrier or an opportunity* (https://ec.europa.eu/digital-single-market/en/blog/multilingualism-digital-age-barrier-or-opportunity).

[88] https://www.europarl.europa.eu/meetdocs/2009_2014/documents/stoa/dv/04ai2_worksh_translation_progr_/04ai2_worksh_translation_progr_en.pdf

[89] https://www.europarl.europa.eu/stoa/en/events/details/language-equality-in-the-digital-age-tow/20161219WKS00342

[90] https://www.europarl.europa.eu/stoa/en/events/details/towards-full-digital-language-equality-i/20220711WKS04301

and opportunities for DLE in the EU.[91] The Workshop was chaired by Mr Jordi Solé (MEP), and the programme included over ten speakers from different parts of Europe for keynotes and a final panel discussion.[92] The central topic was the goal of achieving DLE in Europe and avoiding the digital extinction that at least 21 European languages are currently facing and the importance of protecting multilingualism. The panel presented the results from the European Language Equality[93] (ELE) project, in particular the strategic agenda and roadmap towards achieving full DLE by 2030 through the ELE Programme.[94] The **ELE Programme** (see Section 6 in this document for more detail) is foreseen to be a shared, long-term, coordinated and collaborative LT funding programme tailored to Europe's needs, demands and values — among others, multilingualism and language equality in general. In terms of sharing, the EU has the role of providing resources for coordinating the programme, for providing shared infrastructures, for maintaining the scientific goals and programme principles etc. On the other hand, 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, which 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 NLU 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 features as well as grounding and other sophisticated features that are 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 NLU, 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.

In parallel, the EP's CULT Committee adopted a resolution on AI in the cultural, creative and educational sector in which multilingual and linguistic diversity is also taken into account.[95], [96] Regarding the latter, the resolution calls for: 1) AI technologies to be regulated and trained in order to ensure non-discrimination, gender equality, pluralism, as well as cultural and linguistic diversity; 2) specific indicators to measure diversity in order to promote European ventures and prevent algorithm-based recommendations that negatively affect the EU's cultural and linguistic diversity; and 3) an ethical framework for the use of AI technologies in EU media that guarantees access to culturally and linguistically diverse content. Such a framework should also address the misuse of AI to disseminate fake news and disinformation.[97] In addition, the EC has recently commissioned a study that explores the opportunities of applying AI technologies in ten domains that belong to the cultural, creative and educational sector. This study aims to inspire creative entrepreneurs as well as policy-makers with concrete use cases and recommendations for the application of AI in ten

[91] https://www.europarl.europa.eu/stoa/en/document/EPRS_ATA(2022)729550
[92] https://epthinktank.eu/2022/11/15/achieving-full-digital-language-equality-in-a-multilingual-european-union/
[93] https://european-language-equality.eu/
[94] https://european-language-equality.eu/agenda/
[95] https://www.europarl.europa.eu/news/en/press-room/20210311IPR99709/ai-technologies-must-prevent-discrimination-and-protect-diversity
[96] https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1663438&t=e&l=en
[97] https://op.europa.eu/en/publication-detail/-/publication/b8722bec-81be-11e9-9f05-01aa75ed71a1

cultural and creative sectors.[98] The study also considers language diversity as an opportunity and a risk since AI in Cultural and Creative Sectors (CCS) is largely language-based (NLP, NLU, speech technologies, language-centric AI).

Finally, the conclusions of the Education, Youth, Culture and Sport Council, held on 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.[99]

A key throughline in these documents and initiatives is the idea that LT must be *made in* Europe *for* Europe. This approach will not only strengthen Europe's place at the pole position of research excellence, but also contribute to future European cross-border and cross-language communication, economic growth and social stability. The past few years have witnessed a flurry of white papers and SRAs offering roadmaps and recommendations for how best to attain the goal. In 2019, the European Language Resource Coordination (ELRC) white paper, *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters*, underscored that the main challenge is a lack of appreciation for the value of language data.[100] To help overcome this perception, the group issued several recommendations aimed at the European and national policy level, including:

- Updating the Open Data Directive (2019/1024/EU) so that it references language data as a high-value data category.[101]
- Conducting of a study on language data to identify and quantify the value of language data for citizens, public administrations and businesses.
- Updating national policies (e. g., Open Data policies, digital agenda or strategies for AI) to explicitly support the sharing of language data and LT.
- Including obligatory (language) data management plans in all relevant national funding policies and calls for proposals if not yet included.
- Conducting national surveys to assess translation practices in public administrations at all levels.

These steps will contribute to the development of an inclusive European digital society, a task for which European LT is essential. However, still others are required. The *Report on the Joint Stakeholder Consultation on Research and Innovation in Web Accessibility and Language Technologies*, for instance, highlights that greater work must be done to develop systems capable of adapting and personalizing digital content according to individual needs, particularly in terms of accessibility and language.[102] Research into sign languages represents one avenue that merits greater attention, also considering that sign languages are increasingly becoming recognised as official national languages. Another aspect to take into account is the accessibility of information in multimodal contexts with respect to formatting and the understanding of content.

Happily, it is evident that the EU is well aware of LT's crucial role in building Europe's digital society and has already begun to dedicate funding and launch initiatives to advance LT

[98] https://digital-strategy.ec.europa.eu/en/library/study-opportunities-and-challenges-artificial-intelligence-ai-technologies-cultural-and-creative

[99] https://www.consilium.europa.eu/en/meetings/eycs/2022/04/04-05/

[100] https://lr-coordination.eu/sites/default/files/Documents/ELRCWhitePaper.pdf

[101] https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

[102] https://ec.europa.eu/digital-single-market/en/news/report-joint-stakeholder-consultation-research-and-innovation-web-accessibility-and-language-0. See also the New European Media's SRIA: https://nem-initiative.org/wp-content/uploads/2020/06/nem-strategic-research-and-innovation-agenda-2020.pdf?x98588

and AI. Research, industry, and the public sector have benefitted from these actions. Two prominent examples include the Horizon 2020 Programme and the Connecting Europe Facility (CEF).[103] LT was embedded in the former within research and innovation in the field of information technologies, content technologies, multilingual internet and AI. Through the latter, MT tools (eTranslation[104]) and tools for the management of thesauri and glossaries have been developed (VocBench).[105]

The *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem* provides an analysis of the EU's LT market (including Norway and Iceland) and the adoption of LT by public administrations, both at the EU and national levels.[106] The report underscores that EU industry is fragmented and that many small players struggle to compete with the global giants that dominate the market. It further notes that European businesses and the public sector have become dependent on these non-European global companies, which have massive amounts of data at their disposal due to both copyright disparities between the EU (explicit permission required by European entities) and the US (fair use copyright exception), as well as intensive use of their popular systems. However, the dependency on American or Chinese systems and the torrent of data flowing out of Europe mask areas in which European initiatives may make real the ideal of LT made in Europe for Europe. Several large international tech companies, by way of example, provide MT services free of charge. EU industry, by contrast, is experienced in navigating through Europe's many languages and European MT developers have successfully deployed services for the public sector through the support of EU-funded programmes. LT made for Europe means harnessing this know-how to support MT for all its languages and create domain-specific and application-specific MT while simultaneously being attentive to security and privacy issues. Moveover, as stated in *My Europe. My language: With language technologies made in the EU*,[107] LT offers opportunities to reduce language barriers across Europe and in the DSM at the intersection of Big Data, AI and HPC. Indeed, the European High Performance Computing Joint Undertaking[108] (EuroHPC JU), a joint initiative between the EU, European countries and private partners, is developing a world-class supercomputing ecosystem in Europe that will include a Language Data Space.[109] The corresponding Language Data Space (LDS) EU project is about to start in January 2023.

The EC has also established public-private partnerships (PPPs) in the area of AI.[110] As detailed by Curry et al. (2021), the Big Data Value PPP created by the EC and the BDVA in 2014 represented a substantial collective effort on the part of the European data community to formulate a set of technical research priorities for Big Data. According to the report, Europe's multilingualism presents a particular challenge when it comes to data:

> Large amounts of data are being made available in a variety of formats ranging from unstructured to semi-structured to structured formats … A great deal of this data is created or converted and further processed as text. Algorithms or machines are not able to process the data sources due to the lack of explicit semantics. In Europe, text-based data resources occur in many different languages, since

[103] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/information-and-communication-technologies; https://ec.europa.eu/digital-single-market/en/connecting-europe-facility; https://ec.europa.eu/digital-single-market/en/language-technologies

[104] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation; https://ec.europa.eu/education/knowledge-centre-interpretation/eu-initiatives-language-technologies_en

[105] https://ec.europa.eu/isa2/solutions/vocbench3_en

[106] https://op.europa.eu/en/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en/format-PDF/source-106906783

[107] https://ec.europa.eu/digital-single-market/en/news/my-europe-my-language-language-technologies-made-eu-brochure

[108] https://eurohpc-ju.europa.eu

[109] https://digital-strategy.ec.europa.eu/en/activities/work-programmes-digital

[110] https://adr-association.eu

> customers and citizens create content in their local language. This multilingualism of data sources means that it is often impossible to align them using existing tools because they are generally available only in the English language. Thus, the seamless aligning of data sources for data analysis or business intelligence applications is hindered by the lack of language support and gaps in the availability of appropriate resources.[111]

The Big Data Value PPP's successor, the Data, AI and Robotics Partnership (formed in 2020 along with BDVA,[112] euRobotics,[113] ELLIS,[114] CLAIRE,[115] and EurAI[116]) expanded on this issue and zeroed in on NLP's importance in its Strategic Research, Innovation and Deployment Agenda:[117] "Natural Language Processing has particular resonance within Europe's multilingual landscape and offers the potential to harmonise human interaction." Unfortunately, although the PPP includes LT experts, research groups and companies as members of some of its involved associations, currently no European LT association or network is represented in the PPP.

The initiative, however, complements the Coordinated Plan on Artificial Intelligence (CPAI) proposed by the European Commission for the period 2021-2027. The plan, which considers AI an area of strategic importance and aims to propel Europe to the forefront in terms of developing and exploiting AI technologies, calls for the EU to provide a minimum one billion euro annual investment in Horizon Europe and Digital Europe, although the objective is to reach twenty billion euros a year between public and private investments.[118] The focus is on four key areas: increasing investment in AI; the availability of data; the promotion of talent; and ensuring security, ethics and trust in AI. Success in these domains leans on the belief that member states must develop and coordinate their own national AI strategies, of which an analysis and comparison is provided in the report *AI Watch: National strategies on Artificial Intelligence: A European perspective in 2019*.[119]

4.2 National and regional initiatives

The perspective that member states should be responsible for their individual AI strategies stems partly from the observation that each country or region is best positioned to address their own particular needs, based on the specific local circumstances. The response by European countries to the CPAI has been largely positive and the number of states with an AI strategy (29 out of 30; only Croatia has no official strategy as of yet) demonstrates its success. Moreover, it is in the national plans that currently exist where many of the initiatives concerning LT and language-centric AI reside, although this is not to say that dedicated LT programmes are widespread in Europe. And in comparison to non-EU national AI initiatives, Europe's member states lag behind when LT is taken into account. Keep in mind that since Canada published the world's first national AI strategy in 2017, more than 30 other countries and regions have published similar documents as of December 2020.[120] Several non-EU nations merit brief consideration here due to the explicit inclusion of NLP in their plans. China's AI strategy, one of the most comprehensive in the world, singles out NLU

[111] https://elements-of-big-data-value.eu/research-priorities-for-big-data-value/#page-content
[112] https://www.bdva.eu
[113] https://www.eu-robotics.net
[114] https://ellis.eu
[115] https://claire-ai.org
[116] https://eurai.org
[117] https://adr-association.eu/wp-content/uploads/2020/09/AI-Data-Robotics-Partnership-SRIDA-V3.0-1.pdf
[118] https://knowledge4policy.ec.europa.eu/ai-watch/coordinated-action-plan-ai_en
[119] https://ec.europa.eu/jrc/en/publication/ai-watch-national-strategies-artificial-intelligence-european-perspective-2019
[120] https://aiindex.stanford.edu/report/

technology as a decisive area to promote in university AI curricula and in its pursuit of AI talent.[121] The UK, which emphasizes a strong partnership between business, academia, and government, created a pilot programme for under-18 year olds to encourage careers in the AI sector, explicitly mentioning NLP. India's approach to AI considers the multilingual reality of the country a means to achieve technological leadership in AI and cites the development of an advanced NLP infrastructure for its languages as a stepping stone in that direction.[122] Finally, the United States emphasizes the crucial role LT plays in AI, and NLU appears as one of the six "Uses for Deployed AI Today" in the National Security Commission on Artificial Intelligence's *Final Report*, published in 2021.[123]

In Europe, only a handful of dedicated national programmes funded projects related to LT before 2018.[124] Instead, financial support for the development of LT was generally provided through generic R&D&I calls in most member states. The Spanish case is one of those notable exceptions. The Spanish government has recently announced a new strategic plan for economic recovery and transformation (PERTE) called "The New Economics of Language".[125] The PERTE is presented as an opportunity to take advantage of the potential of Spanish and co-official languages for economic growth and international competitiveness in areas such as AI, translation, learning, cultural dissemination, audiovisual production, research and science. It has a budget of 1.1 billion euros in public funds and aims to mobilize another billion in private investment. Additionally, following the lines of the Spanish Plan for the Advancement of LT,[126] several regional governments have also launched LT initiatives, including AINA (Catalonia),[127] Nós (Galicia)[128] and GAITU (the Basque Country).[129]

At the European level, LT received better support through calls in various programmes: FP7, H2020, CEF Telecom, CIP ICT-PSP, EUREKA and EUROSTARS, among others. However, in these too, most funding for LT projects gradually reduced. If we compare these findings to those presented by Rehm et al. (2020b), we observe a slight increase in the number of language-centric AI initiatives over the next couple of years (these results are updated in Table 1 with the most current reports on national initiatives in Europe).[130] It is noteworthy that only 12 European countries out of the 30 studied explicitly consider LT within their national policy initiatives. This is significant because the successful development of the next generation of innovative AI technology relies on setting aside funding exclusively for LT. The same holds true for European countries that hope to incorporate LT-based AI applications, such as interactive dialogue systems and personal virtual assistants, into public services.[131]

In summary, Europe's multilingual nature is also one of the main obstacles to a truly connected, cross-lingual communication and information space. Moreover, while language

[121] Zhang et al. (2021)

[122] *AI in India: A Policy Agenda*. The report also highlights natural language voice recognition as a way to to account for the diversity in languages and digital skills in the Indian context and recommends the creation of annotated data sets for their languages to add incremental value to existing services ranging from e-commerce to agricultur.

[123] https://www.nscai.gov/2021-final-report. See also, the *American AI Initiative*.

[124] Spanish *Plan for the Advancement of Language Technology*: https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/tecnologias-del-lenguaje-en-Europa.aspx

[125] https://planderecuperacion.gob.es/como-acceder-a-los-fondos/pertes/perte-nueva-economia-de-la-lengua

[126] https://plantl.mineco.gob.es/Paginas/index.aspx

[127] https://politiquesdigitals.gencat.cat/ca/tic/aina-el-projecte-per-garantir-el-catala-en-lera-digital/

[128] https://www.xunta.gal/hemeroteca/-/nova/134792/xunta-usc-ponen-marcha-lsquo-proxecto-nosrsquo-que-permitira-incorporar-galego

[129] 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

[130] Rehm et al. (2020b). According to the authors, only four of the 30 surveyed countries do not have some level of LT funding. Four countries have programmes dedicated to LT (Denmark, Estonia, Iceland, Spain), six provide funding for LT-related topics through AI (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. See also: Rehm et al. (2014a, 2016, 2020a, 2021)

[131] https://digital-strategy.ec.europa.eu/en/news/new-report-looks-ai-national-strategies-progress-and-future-steps

	LT-related funding			Artificial Intelligence	
	None at all	Some funding	Dedicated LT programme	AI strategy	LT funding through AI
Austria	X			X	
Belgium		X		D	X
Bulgaria		X		X	
Croatia	X				
Cyprus				X	
Czechia		X		X	
Denmark			X	X	X
Estonia			X	X	X
Finland		X		X	
France		X		X	X
Germany		X		X	X
Greece		X		D	
Hungary		X		X	
Iceland			X	X	
Ireland		X		X	
Italy		X		X	
Latvia		X		X	
Lithuania		X		X	
Luxembourg		X		X	
Malta		X		X	X
Netherlands		X		X	
Norway		X		X	
Poland		X		X	
Portugal		X		X	
Romania		X		D	
Serbia	X			X	
Slovakia	X			X	
Slovenia		X		X	
Spain			X	X	
Sweden		X		X	

Table 1: Overview of the Language Technology funding situation in Europe (2019/2021), extracted from Rehm et al. (2020b) and updated with the newest AI strategies. D stands for draft documents.

diversity is at the core of European identity, many of our languages are in danger of digital extinction because they are not sufficiently supported through LT (Moseley, 2010; Rehm and Uszkoreit, 2012; STOA, 2017; European Parliament, 2018).[132] Sophisticated multilingual, cross-lingual and monolingual LT 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 faced increased competition from other continents, especially with respect to breakthroughs in AI. These scientific advancements have led to global commercial successes, from which especially the respective regions benefit. As a consequence, many European scientists, including young high-potential researchers, are leaving Europe to continue their work abroad. Europe must invest in retaining and attracting these researchers. Our continent is in need of powerful LT *made in* Europe *for* all European citizens, tailored to our unique cultures, societies and economic requirements so that a linguistically fragmented Europe may become a truly unified and inclusive one. This ambitious but worthy effort involves supporting its 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, benefiting European citizens, European industry and European society.

5 SWOT Analysis

Taking into account all the reports, documents and national and international initiatives that have been analyzed in detail, 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 DLE a reality in Europe by 2030.

5.1 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 MT 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 data 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).

[132] http://www.unesco.org/languages-atlas/index.php?hl=en&page=atlasmap

5.2 Weaknesses

- Deep learning LT and large pre-trained language models have shortcomings and limitations. 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 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.
- 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, 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. AI is often perceived in a polarized fashion as either “magical” technology that can solve any problem, or as a threat to jobs and workers to be replaced by machines.
- 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 (far) fewer LT resources and tools including language resources, annotated corpora, 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.
- The investment in AI does not reflect the real importance of LT.
- 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.

5.3 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 neighbouring 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 organize 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 in the AI/LT-educated workforce.
- Increasing awareness about the possibilities of AI and LT and the necessity to invest and coordinate efforts.

- 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.
- Ensure openness of infrastructures for data and technologies.

5.4 Threats

- As reported by the META-NET White Paper series, at least 21 European languages are in danger of digital extinction, thwarting the fundamental concept of the languages of Europe being equal.
- Development of non-explainable techniques and deep learning models without any commonsense knowledge, with social biases, containing personal and private data, with a very heavy impact on carbon footprint, etc.
- 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 retain in, or attract to, the EU researchers and workers skilled in LT and AI.
- 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.

6 Recommendations

The plan and vision described in this document are compatible with current EU policy, needs and demands, and it is mission-critical to successfully address them. Insufficient investment in the underdeveloped areas of LT and language-centric AI will result in the digital extinction of languages with only global languages spoken by a large numbers of speakers prevailing. Although the overall EU LT community is highly important, without adequate support the global LT/NLP market will be dominated by the US and a few Asian countries, and the European LT community will be pushed aside.

The main requirement of the future large-scale ELE Programme is a collaboration between the EU/EC and all participating countries and regions. Moreover, funding and further investment are needed at all levels. Funding at 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 at the European level is needed because language communities are still too fragmented and small. Further effort should also be geared towards establishing 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. A description of the desirable implementation of these important aspects is given, among others, in the series of XX ELE 1 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 those discussed 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 past 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 collaboration 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 *Deep NLU*, there is a great need to further finance and investigate fields such as cognitive, symbolic and pattern-based AI.

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, as well as natural language understanding and processing tasks such as machine reading, text analysis, MT, chatbots, virtual assistants or summarisation.

Finally, we outline a number of concrete implementation recommendations to help bring about the twin goals of *Deep NLU* and *Digital Language Equality* in Europe by 2030.

6.1 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[133] 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.

6.2 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.

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

6.3 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), making efficient and effective green LT (i. e. technologies with low-demand computational footprint) a recognised priority.
- 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.

6.4 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, in order to 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.

6.5 Research recommendations

6.5.1 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 efficient and effective 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.

6.5.2 Machine Translation

- To develop direct and near-real-time speech-to-speech MT and adaptive MT, where the system learns from user input.
- To develop low-resource MT, by deepening research on embedding projection and structural organisation of embeddings to understand 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 go 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.

6.5.3 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.

6.5.4 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.

6.6 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 details of 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).

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. Deliverable D1.2 Report on the state of the art in Language Technology and Language-centric AI, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

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. Deliverable D3.1 Report on existing strategic documents and projects in LT/AI, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/06/ELE___Deliverable_D3_1__second_revision_2.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

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.

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.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.376. URL https://aclanthology.org/2022.acl-long.376.

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.

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.

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.

Michael Chui, Martin Harryson, James Manyika, Roger Roberts, Rita Chung, Ashley van Heteren, and Pieter Nel. Notes from the ai frontier: Applying ai for social good. *McKinsey Global Institute*, 2018.

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.

Edward Curry, Andreas Metzger, Sonja Zillner, Jean-Christophe Pazzaglia, and Ana García Robles. The elements of big data value: Foundations of the research and innovation ecosystem, 2021.

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, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.417.

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.

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.

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.

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.

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

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.

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.

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.

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, 2013. 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.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*, 2021a.

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

Christopher Moseley. Atlas of the world's languages in danger, 3rd edn., 2010. URL Onlineversion:http://www.unesco.org/culture/en/endangeredlanguages/atlas.

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

Thomas Padilla. Responsible operations: Data science, machine learning, and ai in libraries. *American Archivist*, 83:483–487, 2020.

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.

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.

Georg Rehm. Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda, 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, editor. *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer, Cham, Switzerland, January 2023.

Georg Rehm and Stefanie Hegele. Language technology for multilingual Europe: An analysis of a large-scale survey regarding challenges, demands, gaps and needs. 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/L18-1519.

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, 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. 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 1517–1524, Reykjavik, Iceland, May 2014a. European Language Resources Association (ELRA).

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 2014b.

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, 2016. 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, 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.

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.

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.

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.

Claudia Soria, Nicoletta Calzolari, Monica Monachini, Valeria Quochi, Núria Bel, Khalid Choukri, Joseph Mariani, Jan Odijk, and Stelios Piperidis. The language resource strategic agenda: the flarenet synthesis of community recommendations. *Language Resources & Evaluation*, (48):753–775, 2014. URL https://doi.org/10.1007/s10579-014-9279-y.

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/.

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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

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.

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.

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.

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.

Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault. The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*, 2021. URL https://arxiv.org/abs/2103.06312.

Appendix

A Documents, reports and initiatives

	Year	Title	LT/AI	Country
1	2010	20-Year Strategy for the Irish Language 2010-2030	LT	Ireland
2	2011	Language Resources for the Future – The Future of Language Resources	LT	Europe
3	2011	The FLaReNet Strategic Language Resource Agenda	LT	Europe
4	2012	Special Eurobarometer 386	LT	EU
5	2012	The Charter of Fundamental Rights of the EU	Any	EU
6	2013	Digital Language Death	LT	International
7	2013	LT 2013: Status and potential of the European Language Technology Markets	LT	Europe
8	2013	META-NET Strategic Research Agenda for Multilingual Europe 2030	LT	Europe
9	2013	Unstructured Data and the 80 Percent Rule	LT	International
10	2014	Connecting Europe Facility in Telecom	AI/LT	EU
11	2014	Horizon 2020	AI/LT	EU
12	2015	Cornish Language Strategy 2015-2025	LT	UK
13	2015	Plan for the Advancement of Language Technology	LT	Spain
14	2015	Strategia per la crescita digitale 2014-2020	AI/LT	Italy
15	2015	Strategic Agenda for the Multilingual Digital Single Market – Technologies for Overcoming Language Barriers towards a truly integrated European Online Market.	LT	Europe
16	2015	Wikidata:Lexicographical data/Development/Proposals/	LT	Germany
17	2016	Apropriarse de las redes para fortalecer la palabra	LT	International
18	2016	General Data Protection Regulation	Any	EU
19	2016	Language as a Data Type and Key Challenge for Big Data. Strategic Research and Innovation Agenda for the Multilingual Digital Single Market. Enabling the Multilingual Digital Single Market through technologies for translating, analysing, processing and curating natural language content.	LT	Europe
20	2016	Opening doors to Universal Access to the Media	AI	Europe
21	2017	A Glimpse into Babel: An Analysis of Multilinguality in Wikidata	LT	UK/Germany
22	2017	Assessment of The State of the EU Language Technology Sector and EU Policy Recommendations	LT	Europe
23	2017	Finland’s Age of Artificial Intelligence	AI	Finland
24	2017	Language equality in the digital age – Towards a Human Language Project	LT	EU
25	2017	Language Technologies for Multilingual Europe: Towards a Human Language Project	LT	Europe
26	2017	Language Technology for Icelandic 2018-2022 – Project Plan	LT	Iceland
27	2017	The Next Generation for Artificial Intelligence Plan	AI	China
28	2017	UNESCO Atlas of the World’s Languages in Danger	LT	International
29	2018	AI for Humanity	AI	France

	Year	Title	LT/AI	Country
30	2018	AI in India: A Policy Agenda	AI	India
31	2018	AIM AT 2030 Artificial Intelligence Mission Austria 2030	AI	Austria
32	2018	AI policy report	AI	France
33	2018	A Mission for Europe: Empowering a Multilingual Continent	LT	Europe
34	2018	Artificial Intelligence: A European Perspective	AI	EU
35	2018	Artificial Intelligence Innovation Action Plan for Institutions of Higher Education	AI	China
36	2018	Artificial Intelligence Strategy	AI	Germany
37	2018	Basque a digital language. DLDP Survey Report	LT	Europe
38	2018	Breton a digital language. DLDP Survey Report	LT	Europe
39	2018	Coordinated Plan on Artificial Intelligence	AI	EU
40	2018	Digital Language Diversity Project Roadmap	LT	Europe
41	2018	Digital Language Survival Kit	LT	Europe
42	2018	Estonian Language Technology 2018-2027	LT	Estonia
43	2018	European Artificial Intelligence (AI) leadership, the path for an integrated vision	AI/LT	EU
44	2018	Government report on information policy and artificial intelligence	AI	Finland
45	2018	Irish language strategy 2019-2023	LT	Ireland
46	2018	Karelian a digital language. DLDP Survey Report	LT	Europe
47	2018	Language Equality in the Digital Age	LT	EU
48	2018	Language Technologies	LT	EU
49	2018	Learning to Generate Wikipedia Summaries for Underserved Languages from Wikidata	LT	International
50	2018	L'intelligenza artificiale al servizio del cittadino	AI	Italy
51	2018	National Approach for Artificial Intelligence	AI	Sweden
52	2018	National Strategy on Artificial Intelligence	AI	India
53	2018	Sardinian a digital language. DLDP Survey Report	LT	Europe
54	2018	Work in the Age of Artificial Intelligence	AI	Finland
55	2019	A comprehensive European industrial policy on artificial intelligence and robotics	LT	EU
56	2019	Action plan for the digital transformation of Slovakia for 2019-2022	AI	Slovakia
57	2019	AI for Belgium	AI	Belgium
58	2019	AI in 2019	AI	USA
59	2019	AI in Education: Challenges and Opportunities for Sustainable Development	AI	International
60	2019	AI in the media and creative industries	AI	Europe
61	2019	AI Portugal 2030	AI	Portugal
62	2019	AI Watch – National strategies on Artificial Intelligence: A European perspective in 2019	AI	EU
63	2019	Artificial Intelligence: a strategic vision for Luxembourg	AI	Luxembourg

	Year	Title	LT/AI	Country
64	2019	Artificial Intelligence Strategy of the Valencian Community	AI	Spain
65	2019	AuroraAI – Towards a human-centric society	AI	Finland
66	2019	Beijing Consensus on Artificial Intelligence and Education	AI	International
67	2019	Digital Wallonia 4 AI	AI	Belgium
68	2019	Dutch Digitalisation Strategy 2.0	AI	Netherlands
69	2019	ELRC White Paper	LT	Europe
70	2019	Estonia's national artificial intelligence strategy 2019-2021	AI	Estonia
71	2019	European legislation on open data	AI/LT	EU
72	2019	Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem	LT	EU
73	2019	Flemish Action Plan AI	AI/LT	Belgium
74	2019	Framework for Developing a National Artificial Intelligence Strategy	WEF	International
75	2019	Glimpses of the future: Data policy, artificial intelligence and robotisation as enablers of wellbeing and economic success in Finland	AI	Finland
76	2019	Language technology for digital humanities: introduction to the special issue	LT	Germany/USA
77	2019	Leading the way into the age of artificial intelligence – Final report of Finland's Artificial Intelligence Programme 2019	AI	Finland
78	2019	L'Estratègia d'Intel·ligència Artificial de Catalunya	AI	Spain
79	2019	Liability for AI and other emerging technologies	AI	EU
80	2019	Lithuanian Artificial Intelligence Strategy: a vision for the future	AI	Lithuania
81	2019	Malta AI Strategy	AI	Malta
82	2019	My Europe. My language. With language technologies made in the EU	LT	EU
83	2019	National Strategy for Artificial Intelligence	AI	Denmark
84	2019	News, Disinformation & Language Intelligence (orientation paper)	AI/LT	Europe
85	2019	Regulating disinformation with artificial intelligence	AI/LT	EU
86	2019	Report on the Joint Stakeholder Consultation on Research and Innovation in Web Accessibility and Language Technologies	LT	EU
87	2019	Spanish RDI Strategy in Artificial Intelligence	AI	Spain
88	2019	Sprogteknologi i verdensklasse (World class language technology)	LT	Denmark
89	2019	Strategic Action Plan for Artificial Intelligence	AI	Netherlands
90	2019	Strategic Research, Innovation and Deployment Agenda for an AI PPP A focal point for collaboration on Artificial Intelligence, Data and Robotics	AI	Europe
91	2019	Strategy for the Development of AI in the Republic of Serbia for the period 2020-2025	AI	Serbia
92	2019	Strategy of the Digital Transformation of Slovakia 2030	AI	Slovakia

	Year	Title	LT/AI	Country
93	2019	Ten recommendations for a co-programmed European partnership in AI	AI	NL/Europe
94	2019	The Changing Nature of Work and Skills in the Digital Age	AI/LT	EU
95	2019	The National Artificial Intelligence Strategy of the Czech Republic	AI	Czech Republic
96	2019	The overall view of artificial intelligence and Finnish competence in the area	AI	Finland
97	2019	White Book on EU Public Administrations Translation Contracts	LT	Europe
98	2019	WIPO Technology Trends 2019: Artificial Intelligence	AI	Switzerland
99	2019	Zero to Digital	LT	USA
100	2020	AI and Gender Bias in Recruitment	AI	EU
101	2020	AINA: Catalan language in the digital age	LT	Spain
102	2020	AI Watch: AI Uptake in Health and Healthcare, 2020	AI	EU
103	2020	AI Watch. Artificial Intelligence in public services. Overview of the use and impact of AI in public services in the EU	AI	EU
104	2020	Architecture for a Multilingual Wikipedia	LT	USA
105	2020	Check before you tech	LT	USA
106	2020	CLAIRE Response to the European Commission White Paper – On Artificial Intelligence – A European Approach to excellence and trust	AI	NL/Europe
107	2020	Concept for the Development of Artificial Intelligence	AI	Bulgaria
108	2020	Cornish Language Operational Plan 2019-2020 End of Year Report	LT	UK
109	2020	Digital transformation guidelines for 2021-2027	AI	Latvia
110	2020	Digital Transformation Strategy 2020-2025	AI	Greece
111	2020	ENIA: Estrategia Nacional de Inteligencia Artificial	AI	Spain
112	2020	Global AI Strategy Landscape – 50 National AI Strategies in 2020	AI	International
113	2020	Guidelines for The Development of the Lithuanian Language in The Digital Media And Progress In Language Technologies For 2021-2027	LT	Lithuania
114	2020	Language Technology Programme for Icelandic 2019-2023	LT	Iceland
115	2020	National AI strategy of Cyprus	AI	Cyprus
116	2020	National AI strategy on Developing Artificial Intelligence Solutions	AI	Latvia
117	2020	National Strategy forArtificial Intelligence	AI	Norway
118	2020	Natural Language Processing (NLP) Market to reach US $41 billion by 2025	LT	International
119	2020	NEM SRIA 2020	AI/LT	Europe
120	2020	Slovenia’s National Programme on AI	AI	Slovenia
121	2020	State language policy guidelines for 2021-2027	LT	Latvia
122	2020	stateof.ai	AI	International
123	2020	State of AI in Finland	AI	Finland

	Year	Title	LT/AI	Country
124	2020	Strategia italiana per l’Intelligenza Artificiale	AI	Italy
125	2020	Strategic guidelines for developing AI-solutions	AI	Finland
126	2020	Strategic Research, Innovation and Deployment Agenda	AI	Europe
127	2020	Strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030	LT	EU
128	2020	The Agenda in Brief. Artificial Intelligence in Mexico: A National Agenda	AI	Mexico
129	2020	THEaiTRE: A theatre play written entirely by machines.	LT	Czech Republic
130	2020	The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe	LT	Europe
131	2020	The road to AI. Investment dynamics inthe European ecosystem. AI Global Index 2019	AI	Europe
132	2020	Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability	AI/LT	Europe
133	2020	Trends and Developments in AI – Challenges to the Intellectual Property Rights Framework	AI/LT	EU
134	2020	Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector	AI	England
135	2020	Using NLG to Bootstrap Missing Wikipedia Articles: A Human-centric Perspective	AI/LT	UK
136	2021	2021 Tech Trends Report	AI	International
137	2021	AI index report	AI	USA
138	2021	AI Strategy for Iceland (Draft)	AI	Iceland
139	2021	Artificial Intelligence Development Policy in Poland beyond 2020	AI	Poland
140	2021	Artificial intelligence in EU enterprises	AI/LT	EU
141	2021	Estrategia para la Transformación Digital de Euskadi 2025	AI/LT	Spain
142	2021	Estrategia Procesamiento del Lenguaje Natural 2020	LT	Spain
143	2021	EU Initiatives in language technologies	LT	EU
144	2021	Final Report	AI	USA
145	2021	Global Natural Language Processing Market to Grow at a CAGR of 18.4% from 2020 to 2028	LT	International
146	2021	Natural Language Processing (NLP) – Global Market Trajectory & Analytics	LT	International
147	2021	Natural Language Processing (NLP) Market Size, Share & Industry Analysis	LT	International
148	2021	New Coordinated Plan on Artificial Intelligence	AI	EU
149	2021	Recovery plan for Europe	AI/LT	EU

Table 2: LT and AI reports, documents and initiatives (March 2021).

	Year	Title	LT/AI	Country
1	1992	European Charter for Regional or Minority Languages	LT	EU
2	2020	Nimdzi Language Technology Atlas 2020	LT	International
3	2021	2021 Strategic Foresight Report	AI	International
4	2021	AI Strategy for Iceland	AI	Iceland
5	2021	Classification of AI Systems	AI	International
6	2021	Deep Learning's Diminishinng. The cost of improvement is becoming unsustainable	AI/LT	International
7	2021	Documenting Large Webtext Corpora:A Case Study on the Colossal Clean Crawled Corpus	LT	International
8	2021	EACL 2021 language diversity panel	LT	International
9	2021	ELISE SRA	AI	EU
10	2021	EU looks to make data sharing easier: Council agrees position on Data Governance Act	LT	EU
11	2021	European citizens' knowledge and attitudes towards science and technology	AI	International
12	2021	Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report	AI	International
13	2021	Irish National AI strategy	AI	Ireland
14	2021	National Strategies on Artificial Intelligence: A European Perspective	AI	EU
15	2021	Report on the SME survey on multilingual websites	LT	International
16	2021	State of implementation of the OECD AI Principles: Insights from national AI policies	AI	International
17	2021	Systematic Inequalities in Language Technology Performance across the World's Languages	LT	International
18	2021	The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies	LT	EU
19	2021	The Elements of Big Data Value	AI	EU
20	2021	The Global NLP Market	LT	International
21	2021	The race to understand the exhilarating, dangerous world of language AI	LT	US
22	2021	State of AI Report 2021	AI	International
23	2021	The Inherent Limitations of GPT-3	LT	International
24	2021	CLARIN Vision and Strategy	LT	EU
25	2021	Access to EuroHPC supercomputers is now open	AI	EU
26	2021	The DIGITAL Europe Programme – Work Programmes	AI/LT	EU

Table 3: LT and AI reports, documents and initiatives (October 2021).

	Year	Title	LT/AI	Country
1	2019	Helsinki Initiative on Multilingualism in Scholarly Communication	LT	International
2	2019	Responsible Operations: Data Science, Machine Learning, and AI in Libraries	AI/LT	International
3	2020	AINA, el projecte per garantir el català en l'era digital	LT	Spain
4	2021	$NLP: How to Spend a Billion Dollars	LT	International
5	2021	2021 was the year of monster AI models	LT	International
6	2021	Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model	LT	International
7	2021	Xunta e USC poñen en marcha o 'Proxecto Nós', que permitirá incorporar o galego ás novas linguaxes dixitais	LT	Spain
8	2021	Feasibility study on a legal framework on AI design, development and application based on CoE standards	AI/LT	International
9	2022	The Power of Natural Language Processing	LT	International
10	2022	A.I. Is Mastering Language. Should We Trust What It Says?	LT	International
11	2022	Language Is The Next Great Frontier In AI	LT	International
12	2022	A Wave Of Billion-Dollar Language AI Startups Is Coming	LT	International
13	2022	NLP in Europe is expected to Reach US$35.1 billion by 2026	LT	EU
14	2022	EUROPEAN LANGUAGE INDUSTRY SURVEY 2022: Trends, expectations and concerns of the European language industry	LT	EU
15	2022	GPT-4 Is Coming Soon. Here's What We Know About It	LT	International
16	2022	What can we expect from GPT-4?	LT	International
17	2022	GPT-3, Foundation Models, and AI Nationalism	LT	International
18	2022	Aligning Language Models to Follow Instructions	LT	International
19	2022	Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance	LT	International
20	2022	A New AI Trend: Chinchilla (70B) Greatly Outperforms GPT-3 (175B) and Gopher (280B)	LT	International
21	2022	A new AI draws delightful and not-so-delightful images	LT	International
22	2022	PERTE Nueva economía de la lengua	LT	Spain
23	2022	Democratizing access to large-scale language models with OPT-175B	LT	International
24	2022	Education, Youth, Culture and Sport Council, 4-5 April 2022	LT	EU
25	2022	THE AI INDEX REPORT Measuring trends in Artificial Intelligence	AI/LT	International
26	2022	OECD Framework for the Classification of AI systems	AI/LT	International
27	2022	Unlocking Zero-Resource Machine Translation to Support New Languages in Google Translate	LT	International
28	2022	'The Game is Over': Google's DeepMind says it is on verge of achieving human-level AI	AI/LT	International

	Year	Title	LT/AI	Country
29	2022	Is NLP innovating faster than other domains of AI	AI/LT	International
30	2022	A Generalist Agent	AI/LT	International

Table 4: LT and AI reports, documents and initiatives (April 2022).

	Year	Title	LT/AI	Country
1	2010	Atlas of the world’s languages in danger	LT	International
2	2020	AI, Data and Robotics Partnership Strategic Research, Innovation and Deployment Agenda (version 3.0)		
3	2021	UNESCO World Atlas of Languages: summary document	LT	International
4	2021	Coordinated Plan on Artificial Intelligence 2021 Review	LT/AI	EU
5	2022	Artificial Intelligence for Social Good in Latin America and the Caribbean: The Regional Landscape and 12 Country Snapshots	AI/LT	International
6	2022	AI Watch, road to the adoption of Artificial Intelligence by the public sector	AI/LT	EU
7	2022	Natural Language Processing Market Size is projected to reach USD 91 Billion by 2030, growing at a CAGR of 27%	LT	International
8	2022	Facilitating the implementation of the European Charter for Regional or Minority Languages through artificial intelligence	LT	International
9	2022	Statement of the Committee of Experts of the European Charter for Regional or Minority Languages on the promotion of regional or minority languages through artificial intelligence	LT	International
10	2022	Study on Opportunities and Challenges of Artificial Intelligence (AI) Technologies for the Cultural and Creative Sectors	AI/LT	EU
11	2022	What if everyone spoke the same language?	LT	EU
12	2022	The Strategic Research and Implementation Agenda for achieving full digital language equality in Europe by 2030	LT	EU
13	2022	Achieving full digital language equality in a multilingual European Union	LT	EU

Table 5: LT and AI reports, documents and initiatives (December 2022).