EUROPEAN LANGUAGE EQUALITY

D3.1

Report on existing strategic documents and projects in LT/AI

Authors	Itziar Aldabe (UPV/EHU), Georg Rehm (DFKI), German Rigau (UPV/EHU) and Andy Way (DCU)
Dissemination level	Public
Date	30-04-2021 – Update: 30-11-2021

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D3.1
Deliverable title	Report on existing strategic documents and projects in LT/AI
Type	Report
Number of pages	48
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 31-03-2021 – Actual: 30-04-2021 – Update: 30-11-2021
Work package	WP3: Development of the Strategic Agenda and Roadmap
Task	Task 3.1 Desk research – landscaping
Authors	Itziar Aldabe (UPV/EHU), Georg Rehm (DFKI), German Rigau (UPV/EHU) and Andy Way (DCU)
Reviewers	Jan Hajič (CUNI), Federico Gaspari (DCU)
EC project officers	Aleksandra Wesolowska (April 2021), Susan Fraser and Miklos Druskoczi (November 2021)
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

ACM	Association for Computing Machinery
AI	Artificial Intelligence
CAGR	Compound Annual Growth Rate
CEF	Connecting Europe Facility
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CRS	Congressional Research Service
CULT	European Parliament's Committee on Culture and Education
DSM	Digital Single Market
EC	European Commission
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
EP	European Parliament
ERIC	European Research Infrastructure Consortium
EU	European Union
EurAI	European Association for Artificial Intelligence
G7	Group of Seven
G20	Group of Twenty
GDPR	General Data Protection Regulation
GPAI	Global Partnership on Artificial Intelligence
GPU	Graphical Processing Units
HAI	Human-Centered Artifical Intelligence
HPC	High-Performance Computing
IPR	Intellectual Property Rights
ITRE	European Parliament's Committee on Industry, Research and Energy
JRC	Joint Research Center
LT	Language Technology
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NEM	New European Media

NLP	Natural Language Processing
NLU	Natural Language Understanding
OECD	Organization for Economic Co-operation and Development
PPP	Public-Private Partnership
R&D&i	Research, Development and Innovation
SME	Small and Medium Enterprises
SRA	Strategic Research Agenda
SRIA	Strategic Research and Innovation Agenda
SRIDA	Strategic Research, Innovation and Deployment Agenda
STOA	Science and Technology Options Assessment
WEF	World Economic Forum
WIPO	World Intellectual Property Organization
UN	United Nations
US	United States
USD	US dollar
UNESCO	United Nations Educational, Scientific and Cultural Organization

Remark on November 2021 Version of this Report

The original Deliverable 3.1 *Report on existing strategic documents and projects in LT/AI*, released in April 2021, has been updated in October 2021. The updated version includes references to more than 20 new relevant documents, reports and initiatives that we found since April 2021, the vast majority of which are from 2021. The original references are listed in Table 2 (p. 37 ff.) and the newly identified ones in Table 3 (p. 42). Additionally, more than ten relevant research papers discovered when compiling Deliverable 1.2 *Report on the state of art in Language Technology and Language-centric AI* have also been included.

1. Introduction

This document reports on the initial desk research phase towards the systematic collection and analysis of the existing international, national and regional Strategic Research Agendas (SRAs), studies, reports and initiatives related to Language Technology (LT) and Artificial Intelligence (AI). Around 180 such documents from within the European Union (EU) and international sources have been reviewed and analysed for this purpose. Thus, other Work Packages of the ELE project can be informed at an early stage about any relevant documents that need to be taken into account. All collected AI and LT documents, reports and initiatives are listed in Appendix A.[1] [2]

Natural language is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and manipulate information. In fact, most of the digital information available is unstructured information in the form of documents (written or spoken) in multiple languages, representing a challenge for any organization that wants to exploit and process its information. In fact, up to 80% of all data is unstructured text data.[3] Most computer systems process only structured data (for example databases with millions of records) because it is non-trivial to process *unstructured digital information* (including written and spoken language). Unstructured language data is subject to multiple interpretations (ambiguity), requires knowledge about the context and the world and it is intrinsically complex to process.

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

LT is concerned with studying and developing systems capable of processing human language. The field has developed, over the years, different methods to make the information contained in written and spoken language explicit or to generate or synthesise written or spoken language. Despite the inherent difficulty of many of the tasks performed, current LT support allows many advanced applications which have been unthinkable only a few years ago. LT is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, text editors, text predictors, automatic translation systems, automatic subtitling, automatic summaries, inclusive technology, etc. Its rapid de-

[1] Links and URLs mentioned in the report last accessed online as of 30th April 2021.
[2] Additional links and URLs mentioned in the report last accessed online as of 30th November 2021.
[3] http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

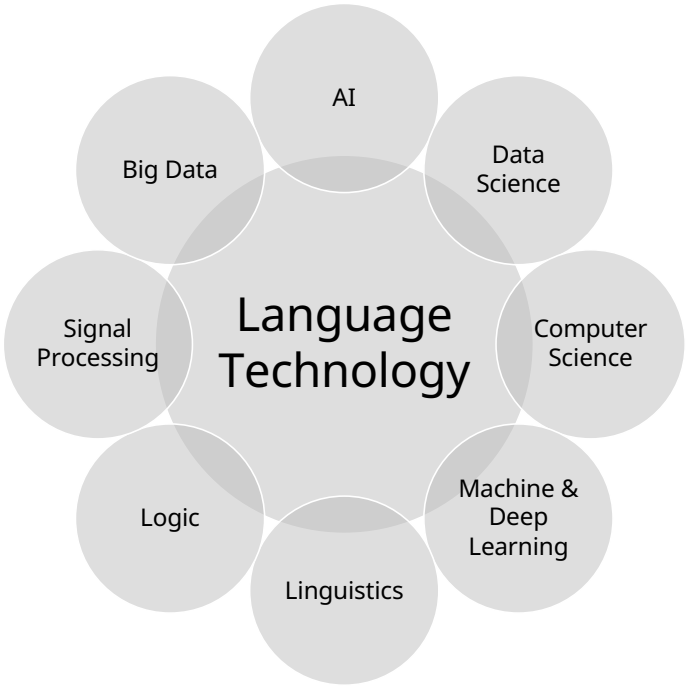

Figure 1: Language Technology as a multidisciplinary field.

velopment in recent years predicts even more encouraging and also exciting results in the near future (see Section 2).

Several recent international reports (see Section 3), prepared between 2018 and 2021, place LT as one of the three core application areas within AI together with Vision and Robotics.[4] Automatic language understanding is perceived as one of the fundamental goals of AI, and, in turn, it is also considered one of its main challenges (Sayers et al., 2021).

LT is at the heart of the software that processes unstructured information and exploits the vast amount of data contained in text, audio and video files including those from the web, social media, etc. Only the proper application of LT will allow processing and understanding, i. e., making sense of these enormous volumes of multilingual written and spoken data in sectors as diverse as health, justice, education, or finance. LT applications such as speech recognition, speech synthesis, textual analysis and machine translation are actually used by hundreds of millions of users on a daily basis.[5] As reflected in the the European, national and regional AI and LT strategies both inside and outside Europe (see Sections 4, 5 and 6), LT is outlined as one of the most relevant technologies for society, as seen by its inclusion in all the prioritized strategic areas for developing research, development and innovation (R&D&i) activities. LT is multidisciplinary in nature since it combines knowledge in computer science (and specifically in AI), mathematics, linguistics and psychology among others. Figure 1 shows some of the most important disciplines involved in LT. This uniqueness must be considered in any public or private initiative in AI that includes LT.

LT is one of the most important AI application areas with a fast growing economic impact. 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

[4] https://oecd.ai/en/classification
[5] https://www.nimdzi.com/nimdzi-language-technology-atlas-2020/

2019,[6] the global NLP market size to grow from USD 10.2 billion in 2019 to USD 26.4 billion by 2024, at a CAGR of 21.0 percent is set during the forecast period 2019-2024.[7] According to another report from the end of 2019,[8] the global NLP market was valued at USD 8.5 billion in 2018, which is expected to reach USD 23.0 billion by 2024, registering a CAGR of 20.0% during the forecast period. A report from 2020 highlights that the global NLP market size stood at USD 8.61 billion in 2018 and is projected to reach USD 80.68 billion at 2026, exhibiting a CAGR of 32.4% during the forecast period.[9] Another report from 2020 estimates the global LT market to reach USD 41 billion by 2025.[10] 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.[11] 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.[12] As a final example, a recent report from 2021 estimates 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 .[13]

In varietate concordia (in English: *united in diversity*[14]) 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.[15] [*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,[16] in addition to migrant languages and various sign languages, spoken by some 50 million people. Figure 2 shows the languages in danger in Europe according to the *UNESCO Atlas of the World's Languages in Danger* (Moseley, 2010).[17] In this map black flags correspond to already extinct languages.

Furthermore, the Charter of Fundamental Rights of the EU under Article 21[18] states that,

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

6 https://www.businesswire.com/news/home/20191230005197/en/Global-Natural-Language-Processing-NLP-Market-Size
7 https://www.analyticsinsight.net/potentials-of-nlp-techniques-industry-implementation-and-global-market-outline/
8 https://www.vynzresearch.com/ict-media/natural-language-processing-nlp-market
9 https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933
10 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
11 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
12 https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market
13 https://www.analyticsinsight.net/the-global-nlp-market-is-predicted-to-reach-us127-26-billion-by-2028/
14 https://europa.eu/european-union/about-eu/symbols/motto_en
15 http://europa.eu/abc/symbols/motto/index_en.htm
16 https://en.m.wikipedia.org/wiki/European_Charter_for_Regional_or_Minority_Languages
17 http://www.unesco.org/languages-atlas/
18 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

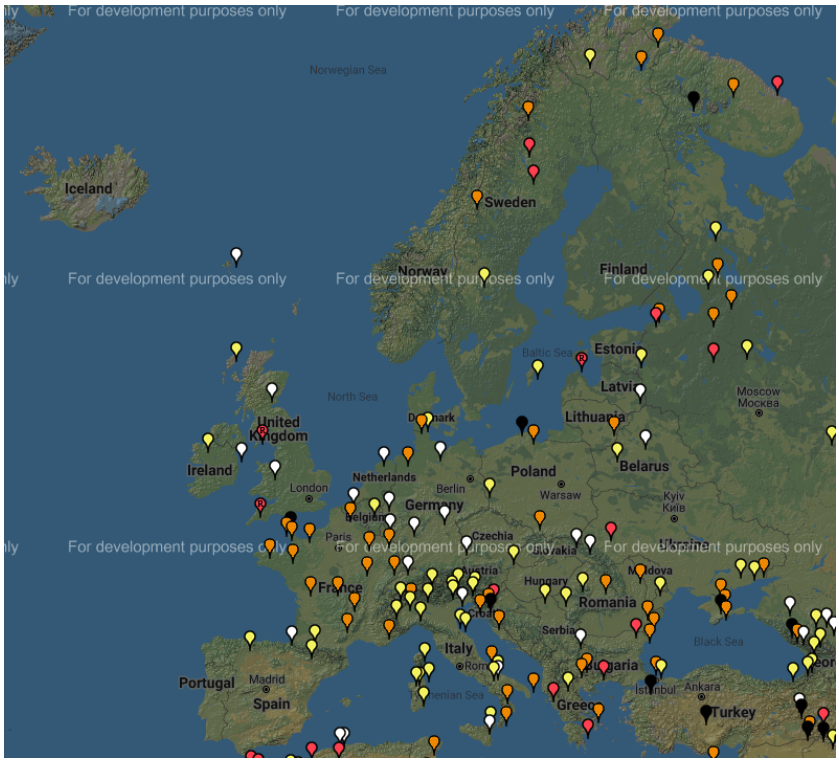

Figure 2: Endangered European languages according to the UNESCO Atlas of the World's Languages in Danger.

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

To help repair the economic and social damage caused by the pandemic, the EU has agreed on a recovery plan to lead the way out of the crisis towards a modern and more sustainable Europe. The EU's long-term budget for 2021-2027, coupled with NextGenerationEU, the temporary instrument designed to boost the recovery, will be the largest stimulus package ever financed through the EU budget. A total of €1.8 trillion will help rebuild a post-COVID-19 Europe.[19] 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.[20]

In this context, the objective of this report is to position language-centric AI, analyzing its strengths and weaknesses, opportunities and threats, showing their unique and multidisciplinary nature in terms of the issues addressed (see Section 7). We also summarize a series of recommendations for the strengthening of LT in Europe (see Section 8).

[19] https://ec.europa.eu/info/strategy/recovery-plan-europe_en
[20] https://digital-strategy.ec.europa.eu/en/activities/digital-programme

2. Language Technology: General Overview

Nowadays, many people use LT every day, especially online – often without even knowing that they do. LT is an important but often invisible ingredient of applications as diverse as, say, search engines, spell-checkers, machine translation systems, recommender systems, virtual assistants, transcription tools, voice synthesizers and many others.

2.1. A very brief historical view

The 1950s marked the beginning of LT as a discipline. In the middle of the 20th century, Alan Turing proposed his famous test, which defined a criterion to determine whether a machine could be considered intelligent (Turing, 1950). A few years later, Noam Chomsky with his generative grammar laid the foundations to formalise, specify and automate linguistic rules (Chomsky, 1957). For a long time, the horizon defined by Turing and the instrument provided by Chomsky influenced the vast majority of NLP research.

The early years of LT were closely linked to MT, a well-defined task, and also relevant from a political and strategic point of view. In the 1950s it was believed that a quality automatic translator would be available soon. After several years of effort, in the mid-1960s 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, in general, of NLP (Pierce and Carroll, 1966). The ALPAC report had a devastating impact on R&D&I funding for the field. From then on, the NLP community turned towards more specific and realistic objectives. The 1970s and 1980s were heavily influenced by Chomsky's ideas, with increasingly complex systems of handwritten rules. At the end of the 1980s, a revolution began which irreversibly changed the field of NLP. This change was driven mainly 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), that would pave the way for subsequent major developments.

Since the 1990s NLP has moved forward, with new resources, tools and applications. Also noteworthy from this period was the effort to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc., of which WordNet (Miller, 1992) is one of the main results. Gradually, data-based systems have been displacing rule-based systems, and today it is difficult to conceive of an NLP system that does not have some component based on ML. In the 2010s we observed a radical technological change in NLP. Collobert et al. (2011) presented a multilayer neural network adjusted by backpropagation which was able to solve various sequential labeling problems. The success of this approach lies in the ability of these networks to learn continuous vector representations of the words (or word embeddings) using unlabelled data (for parameter initialisation) and using labelled data (for fine-tuning the parameters) to solve the task at hand. Word embeddings have played a very relevant role in recent years as they allow the incorporation of pretrained external *knowledge* in the neural architecture (Mikolov et al., 2013; Pennington et al., 2014; Mikolov et al., 2018).

The availability of large volumes of unannotated texts together with the progress in self-supervised Machine Learning and the development of high-performance hardware (in the form of Graphical Processing Units, GPUs) enabled the development of very effective deep learning systems across a range of application areas.

2.2. The Deep Learning era

In recent years, the LT community has witnessed the emergence of powerful new deep learning techniques and tools that are revolutionizing the approach to LT tasks. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of text data. For instance, the *AI Index Report 2021*[21] highlights the rapid progress in NLP, vision and robotics thanks to deep learning and deep reinforcement learning techniques. In fact, the *Artificial Intelligence: A European Perspective* report[22] establishes that the success in these areas of AI has been possible because of the confluence of four different research trends: 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).

As a result, various IT enterprises have started deploying large pretrained neural language models in production. Google and Microsoft have integrated them in their search engines and companies such as OpenAI have also been developing very large language models. Compared to the previous state of the art, the results are so good that systems are claimed to obtain human-level performance in laboratory benchmarks when testing some difficult English language understanding tasks. However, those systems are not robust enough, very sensitive to phrasing and typos, perform inconsistently (when they are faced with similar input), etc. (Ribeiro et al., 2018, 2019). Additionally, existing laboratory benchmarks and datasets also have a number of inherent and severe problems (Caswell et al., 2021). For instance, the ten most cited AI datasets are riddled with label errors, which is likely to distort our understanding of the field's progress (Northcutt et al., 2021).

Forecasting the future of LT and language-centric AI is a challenge. A few years ago, few would have predicted the recent breakthroughs that have resulted in systems that can translate without parallel corpora (Artetxe et al., 2019), create image captions (Hossain et al., 2019), generate full text claimed to be almost indistinguishable from human prose (Brown et al., 2020), generate theatre play scripts (Rosa et al., 2020), create pictures from textual descriptions[23] (Ramesh et al., 2021) and systems able to deal with 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, however, safe to predict that even more advances will be achieved by using pretrained language models. For instance, GPT-3 (Brown et al., 2020), one of the largest dense language models, can be fine-tuned for an excellent performance on specific, narrow tasks with very few examples. GPT-3 has 175 billion parameters and was trained on 570 gigabytes of text, with a cost estimated at more than four million USD.[24] In comparison, its predecessor, GPT-2, was over 100 times smaller, at 1.5 billion parameters. This increase in scale leads to surprising behaviour: GPT-3 is able to perform tasks it was not explicitly trained on with zero to few training examples (referred to as zero-shot and few-shot learning, respectively). This behaviour was mostly absent in the much smaller GPT-2 model. Furthermore, for some tasks (but not all), GPT-3 outperforms state-of-the-art models explicitly trained to solve those tasks with far more training examples. Furthermore, recent work has shown 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

[21] https://aiindex.stanford.edu/report/
[22] https://ec.europa.eu/jrc/en/publication/artificial-intelligence-european-perspective
[23] https://openai.com/blog/dall-e/
[24] https://lambdalabs.com/blog/demystifying-gpt-3/

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 impressive that these models 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. Despite their impressive capabilities, large pretrained language models do come with some drawbacks. For instance, GPT-3 still has serious weaknesses and sometimes makes very silly mistakes.[25] Moreover, their effectiveness across so many tasks demands caution, as their defects are inherited by all the adapted models downstream. Currently we have no clear understanding of how they work, when they fail, and what emergent properties they present. To tackle these questions, much critical interdisciplinary collaboration and research are needed. Thus, some authors call these models *foundation models* to underscore their critically central yet incomplete character (Bommasani et al., 2021). According to Dodge et al. (2021), one of the largest NLP datasets available has been extensively "filtered" to remove Black and Hispanic authors, material related to LGBTQ identities, as well as source data that deals with a number of other minority identities.[26] In short, they can generate racist, sexist, and otherwise biased text. Furthermore, they can generate unpredictable and factually inaccurate text or even recreate private information.[27] Combining large language models with symbolic approaches (knowledge bases, knowledge graphs), which are often used in large enterprises because they can be easily edited by human experts, is a non-trivial challenge. Techniques for controlling and steering such outputs to better align with human values are nascent but promising. These models are also very expensive to train, which means that only a limited number of organisations with abundant resources in terms of funding, computing capabilities, LT experts and data can currently afford to develop and deploy such models. A growing concern is that due to unequal access to computing power, only certain firms and elite universities have advantages in modern AI research (Ahmed and Wahed, 2020).

Moreover, computing large pretrained models also comes with a very large carbon footprint.[28] Strubell et al. (2019) recently benchmarked model training and development costs in financial terms and estimated carbon dioxide emissions. While the average human is responsible for an estimated five tons of carbon dioxide per year,[29] the authors trained a big neural architecture and estimated that the training process emitted 284 tons of carbon dioxide. Finally, such language models have an unusually large number of uses, from chatbots to summarization, from computer code generation to search or translation. Future users are likely to discover more applications, and use positively (such as knowledge acquisition from electronic health records) and negatively (such as generating deep fakes), making it difficult to identify and forecast their impact on society. As argued by Bender et al. (2021), it is important to understand the limitations of large pretrained language models, which they call "stochastic parrots" and put their success in context.

Given the role of LT in everyone's daily lives, many LT practitioners are particularly concerned by language diversity in LT research.[30] For instance, Sayers et al. (2021) emphasise issues of inequality and several groups who will be disadvantaged. Important questions concerning security and privacy will accompany new LT. Looking ahead, they see many intriguing opportunities and new capabilities, but also a range of other uncertainties and inequalities. Joshi et al. (2020) examine the relation between types of languages, resources and their representation in NLP conferences over time. As expected, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving LT field. Blasi et al.

[25] https://lastweekin.ai/p/the-inherent-limitations-of-gpt-3
[26] https://www.unite.ai/minority-voices-filtered-out-of-google-natural-language-processing-models/
[27] https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html
[28] https://spectrum.ieee.org/deep-learning-computational-cost
[29] https://ourworldindata.org/co2-emissions
[30] https://gitlab.com/ceramisch/eacl21diversity/-/wikis/EACL-2021-language-diversity-panel

(2021) study the systematic inequalities in LT across the world's languages. After English, only a handful of Western European languages dominate the field – in particular German, French and Spanish – as well as even fewer non-Indo-European languages, primarily Chinese, Japanese and Arabic. This investigation suggests that it is the economic status of the users of a language (rather than the sheer demographic demand) that drives the development of LT.

In summary, despite claims of human parity in many of the LT tasks, Natural Language Understanding (NLU) is still an *open research problem* far from being solved since all current approaches have *severe* limitations. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pretrained language models, prompt learning and self-supervised systems opens up the way to leverage LT for less developed languages. For the first time, a single multilingual model has outperformed the best specially trained bilingual models on news translations. That is, a single multilingual model provided the best translations for both low- and high-resource languages, showing that the multilingual approach is indeed the future of MT (Tran et al., 2021). However, the development of these new LT systems would not be possible without sufficient resources (experts, data, computing facilities, etc.) as well as the creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application. Focusing on state-of-the-art results exclusively with the help of leaderboards without encouraging deeper understanding of the mechanisms by which they are achieved can generate misleading conclusions, and direct resources away from efforts that would facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication, to create transparent digital language equality in Europe in all aspects of society, from government to businesses to the citizens.

3. Language Technology in International Organizations

AI capabilities are rapidly evolving. AI has become one of the most transformative technologies of the 21st century.[31] In recent years, due to the growing interest in AI at a global political, scientific and social level, several international organizations have developed a number of different reports, discussions and initiatives.

3.1. Reports from International Organizations

The Organisation for Economic Co-operation and Development[32] (OECD) studied the impact of AI on the economy and society in the book *Artificial Intelligence in Society*.[33] The book maps the impact of AI technologies and applications and their policy implications, presenting evidence and policy options. It is also intended to help coordination and consistency with discussions in other international fora, notably the G7, the G20, the EU and the UN. The OECD's 2021 report, *State of the implementation of the OECD AI Principles: Insights from national AI policies* based on national AI policies, adds to the aforementioned study by identifying challenges and good practices for the implementation of five policy recommendations to governments contained in its OECD AI Principles.[34] The five recommendations to governments are:

1. Invest in AI R&D;

[31] https://www.holoniq.com/notes/50-national-ai-strategies-the-2020-ai-strategy-landscape/
[32] https://www.oecd.org
[33] https://doi.org/10.1787/eedfee77-en
[34] https://doi.org/10.1787/1cd40c44-en

2. Foster a digital ecosystem for AI;
3. Shape an enabling policy environment for AI;
4. Build human capacity and preparing for labour market transformation; and
5. Foster international co-operation for trustworthy AI.

The report also gives practical advice for implementing the OECD AI Principles throughout each phase of the AI policy cycle.

The World Economic Forum[35] (WEF) developed a framework to guide governments that are yet to develop national AI strategies or which are in the process of developing them. The framework helps to ask the right questions, follow the best practices, identify and involve the right stakeholders in the process and create the right set of outcome indicators.[36]

The World Intellectual Property Organization[37] (WIPO) *Technology Trends 2019 Artificial Intelligence* report[38] offers unique insights into trends in AI techniques, AI applications (such as NLP, speech processing and computer vision) and AI application fields (i. e., those industries and sectors in which AI is being put into practice). One of the most striking findings of the report is that 50 percent of all AI patents have been published in just the last five years, a remarkable illustration of how rapidly innovation is advancing in this field. Throughout the report, AI technologies are analyzed using a scheme based on the Association for Computing Machinery (ACM) Computing Classification Scheme, which has been developed over the past 50 years. As it was last updated in 2012, the scheme has been adapted to take account of recent technological developments. It comprises three main categories:

- AI techniques: advanced forms of statistical and mathematical models, such as machine learning, deep learning, fuzzy logic and expert systems, allowing the computation of tasks typically performed by humans; different AI techniques may be used as a means to implement different AI functions.
- AI functional applications: functions such as NLP or speech processing, computer vision or robotics which can be realized using one or more AI techniques.
- AI application fields: different fields, areas or disciplines where AI functional applications can be deployed, such as transportation, agriculture or life and medical sciences.

Following this scheme, LT (including NLP and speech processing) is one of the most important functional applications of AI. In fact, according to WIPO's trends in AI, two top areas in functional applications are NLP (14% of all AI-related patents) and speech processing (13%).

Another report from 2020, OECD's *The Digitalisation of Science, Technology and Innovation*,[39] stresses that policies need to ensure that digital technologies are developed to respond to societal challenges. There are many examples of AI applications that tackle these challenges and some are based on NLP techniques. For example, it is possible to identify victims of sexual exploitation on the internet based on face detection, social network analysis and NLP (Chui et al., 2018). This report also emphasises that technological advances in NLP are opening new analytical possibilities and in some countries, researchers and policy makers have started to experiment with NLP. They are using it to track emerging research topics and technologies and to support R&D&I decisions and investments. For instance, in Spain, the Corpus Viewer tool, developed by the State Secretariat for Information Society and Digital

[35] https://www.weforum.org
[36] https://www.weforum.org/whitepapers/a-framework-for-developing-a-national-artificial-intelligence-strategy
[37] https://www.wipo.int
[38] https://www.wipo.int/publications/en/details.jsp?id=4386
[39] https://www.oecd-ilibrary.org/science-and-technology/the-digitalisation-of-science-technology-and-innovation_b9e4a2c0-en

Agenda, processes and analyses large volumes of textual information using NLP. Policy makers use these results to monitor and evaluate public programmes, and to formulate science and innovation policy initiatives.

The UNESCO[40] *Beijing Consensus on Artificial Intelligence and Education*[41] recommends that governments and other stakeholders, in accordance with their legislation and public policies, consider implementing actions in response to the education-related opportunities and challenges presented by AI. In particular, people need to be mindful of the multidisciplinary nature of AI and its impacts, and ensure that AI tools in teaching and learning enable the effective inclusion of students with learning impairments or disabilities and those studying in a language other than their mother tongue. Moreover, LT can also be very helpful in the educational sector, e. g., the UNESCO *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development* report[42] mentions that one of the biggest breakthroughs so far in China has been a system that is able to correct student essays with AI.

The *State of AI Report* analyses the most interesting developments in AI.[43] The report is produced by UK AI investors aiming to trigger an informed conversation about the state of AI and its implications for the future. The report considers the following key dimensions: research, talent, industry, politics and predictions. It highlights that a new generation of language models are unlocking new NLP use-cases and that huge models, large companies and massive training costs dominate the hottest area of AI today: NLP. The last version of this report from 2021 highlights that Transformers have emerged as a general purpose architecture for ML, beating the state of the art in many domains including NLP, computer vision, and even protein structure prediction. Additionally, they also remark the emergence of large language models.[44]

3.2. Reports from the United States

Recently, the Institute for Human-Centered AI (HAI) at Stanford University published its *AI Index 2021* report (Zhang et al., 2021).[45] Compared to previous reports it contains more data and analysis on technical performance, diversity, and ethics and is organized 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. The Research and Development chapter covers the growth of research papers and conferences over time and by region. Technical Performance tracks AI accuracy on several benchmarks in computer vision, NLP, and molecular biology. The Economy chapter focuses on trends in jobs and investment by country, while the AI Education chapter looks at university course offerings and PhD graduates in AI; a key takeaway is that in North America, 65% of the new PhDs chose jobs in industry over academia, compared to 44.4% the previous year. The Ethical Challenges chapter notes that the team was "surprised to discover how little data there is on this topic," and in particular calls out a lack of benchmarks. The Diversity chapter also cites a lack of publicly available data, but does point out that the "AI workforce remains predominantly male." The final chapter, AI Policy and National Strategies, shows trends in various national AI strategies and international collaborations on AI. The Institute has also updated its Global Vibrancy Tool for comparing up to 26 countries across 22 metrics.[46] The metrics measure performance on various research and development, economic, and inclusion factors, such as number of conference papers, number of patents, and investment. The

[40] https://en.unesco.org
[41] https://unesdoc.unesco.org/ark:/48223/pf0000368303
[42] https://unesdoc.unesco.org/ark:/48223/pf0000366994
[43] https://www.stateof.ai
[44] https://www.stateof.ai/2021-report-launch.html
[45] https://aiindex.stanford.edu/report/
[46] https://aiindex.stanford.edu/vibrancy/

tool can show an overall index for the full list of countries, or detailed metrics for a single country, and contains data from the year 2015 up to 2020. The tool covers 13 European countries, i. e., Switzerland, UK, Finland, Germany, France, Sweden, Ireland, Netherlands, Denmark, Norway, Italy, Spain and Belgium. The ranking is led by the United States, Singapore, Switzerland, China, South Korea, India, Israel, Australia, Canada and the United Kingdom. None of the top 10 countries in the ranking is an EU Member State.

A major new report on the state of AI has just been released. The report *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report* comes out of the AI100 project,[47] which aims to study and anticipate the effects of AI as it ripples out through our lives over the course of the next 100 years. AI100 produces a new report every five years. The first report was published in 2016, and this is the second. Members of the writing panel come from across the world, with backgrounds in computer science, engineering, law, political science, policy, sociology and economics. The report highlights the remarkable progress made in AI over the past five years. AI is leaving the laboratory and entering our lives, having a "real-world impact on people, institutions, and culture" (page 71). In particular, the report highlights that the ability of computer programs to perform sophisticated language- and image-processing tasks has advanced significantly. According to this report, a greater investment of time and resources is required to meet the challenges posed by the rapidly evolving technologies of AI and associated fields. In addition to regulation, governments also need to educate. In an AI-enabled world, our citizens, from the youngest to the oldest, need to be literate in these new digital technologies.

The National Security Commission on AI has recently issued a 750-page report that says that the US, which once led in AI, now has just a few years' lead on China and risks being overtaken unless the government steps in.[48] The 16 chapters explain the steps the US must take to responsibly use AI for national security and defense, defend against AI threats, and promote AI innovation. In the report, NLU appears as one of the six *Uses for Deployed AI Today*. The 15-member commission calls for a USD 40 billion investment to expand and democratize AI research and development a *modest down payment* for future breakthroughs, and encourages an attitude toward considering this investment in innovation from policy makers. The group envisions hundreds of billions of dollars of spending on AI by the federal government in the coming years.

According to the 14th Annual Edition 2021 Tech Trends Report on AI[49] from The Future Today Institute[50] "Natural language processing is an area experiencing high interest, investment, and growth". They also forecast that NLP algorithms, typically used for text, words, and sentences, will be used to interpret genetic changes in viruses and that one of the datasets which measures AI English language ability will likely be surpassed by the end of 2021.

The Global AI Talent Tracker report from Macro Polo think tank of the Paulson Institute provides an analysis of the global balance and the flow of top AI scientists.[51] According to this analysis, the US lead in AI is built on attracting international talent, with more than two-thirds of the top-tier AI researchers working in the US having received undergraduate degrees in other countries. In particular, although 18% of the top-tier AI researchers come from Europe, only 10% of them work in Europe.

A major fear surrounding AI is whether or not people are at risk of losing their jobs to AI technologies. The study of Blumberg Capital of 1,000 American adults found that about half are prepared to accept new technology, while the other half are frightened it will take their jobs away.[52] One surprising finding: most individuals (72%) understand that AI is proposed

[47] https://ai100.stanford.edu
[48] https://www.nscai.gov/2021-final-report
[49] https://2021techtrends.com/AI-Trends
[50] https://futuretodayinstitute.com
[51] https://macropolo.org/digital-projects/the-global-ai-talent-tracker/
[52] https://blumbergcapital.com/ai-in-2019/

to remove the exhausting, dull parts of what they do, freeing them to concentrate on more creative tasks. All things considered, 81% are so fearful of being supplanted that they are reluctant to surrender their boring work to an algorithm.

3.3. Reports from the European Union

The second annual Strategic Foresight Report, *The EU's capacity and freedom to act*,[53] presents a forward-looking and multidisciplinary perspective on important trends affecting the EU's capacity and freedom to act in the coming decades. According to this report, the EU's capabilities in artificial intelligence, big data and robotics are similar to Japan's, but it needs to catch up with leaders: the USA and China. The report identifies 10 strategic areas to ensure the EU's freedom and capacity to act in the coming decades. The third area is "Strengthening capacity in data management, artificial intelligence and cutting edge technologies" in order to ensure digital sovereignty. To accomplish this, it encourages stakeholders to promote values via financing, developing and producing the next generation tech, and building capacity to store, extract and process data.

The last roadmap from the European Strategy Forum on Research Infrastructures (ESFRI) (ESFRI-Roadmap, 2018) includes Big Data technology as one of the emerging drivers of the landscape analysis. According to this analysis, LT is a core Big Data technology since the growth in the volume and variety of data is mostly due to the accumulation of unstructured text data. This is why research infrastructures in LT are indispensable in breaking new ground. Regarding LT, the ESFRI Landmark CLARIN ERIC (Common Language Resources and Technology Infrastructure) offers interoperable access to language resources and technologies for researchers in the humanities and social sciences.[54] The CLARIN vision is that "all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on on-line environment for the support of researchers in the humanities and social sciences".[55] Unfortunately, not all EU Member States are official members of CLARIN[56] (i. e. Ireland, Luxembourg, Malta, Slovakia and Spain are not CLARIN members) and some of them just participate in CLARIN as observers (i. e., France). Moreover, as the research funding agencies are providing unbalanced resources to the different Member States, the European languages are not equally supported by CLARIN (de Jong et al., 2020).

Regarding AI, various documents have been published recently by the European institutions: European AI leadership, the path for an integrated vision,[57] the Strategy on AI,[58] the Ethics Guidelines for the Trustworthy AI,[59] Liability for AI and other emerging technologies,[60] the White Paper on AI,[61] and the Coordinated Plan on AI.[62] They all agree that AI is an area of strategic importance and key driver of economic development and that it can provide solutions to many societal challenges. In fact, Europe is strong in core AI systems but unable to make full use of its potential in industrial applications. The socio-economic, legal and ethical impact of AI has to be carefully addressed. For instance, the Joint Research Center (JRC) Science for Policy Report *The Changing Nature of Work and Skills in the Digital Age*[63]

[53] https://ec.europa.eu/info/strategy/strategic-planning/strategic-foresight/2021-strategic-foresight-report_en
[54] http://www.clarin.eu
[55] https://www.clarin.eu/content/vision-and-strategy
[56] In September 2021 Belgium joined CLARIN. https://www.clarin.eu/news/belgium-joins-clarin-eric-member
[57] https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2018)626074
[58] https://ec.europa.eu/digital-single-market/en/artificial-intelligence#Building-Trust-in-Human-Centric-Artificial-Intelligence
[59] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines
[60] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199
[61] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[62] https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence
[63] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/changing-nature-work-and-skills-digital-age

observes that new jobs related to the development and maintenance of AI technologies and big data infrastructures are among those expected to grow, whereas the jobs that are most exposed to automation appear to be those that require relatively low levels of formal education, not involving complex social interaction or predominantly requiring routine manual tasks. However, digital technologies do not simply create and destroy jobs, they also change what people do on the job, and how they do it.

The Member States and the European Commission (EC) agreed to work together to stay at the forefront of AI. If applied in an ecosystem of excellence and trust, European AI can be globally competitive while respecting European values. High-quality data is 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. Thus the 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.[64] However, as 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/GDPR).[65] While the GDPR/Copyright regulations support the privacy and rights of European citizens, it is a major barrier to the access and re-use of language resources, in competition with countries that adopted the "fair use" doctrine, such as the US, Japan or Korea. Fortunately, the EU is working to strengthen various data-sharing mechanisms. Recently, the Member States agreed on a negotiating mandate on a proposal for a Data Governance Act (DGA).[66] The Data Governance Act is 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 used to power applications and advanced solutions in artificial intelligence, personalised medicine, green mobility, smart manufacturing and numerous other areas.

Very recently, the EC presented a New Coordinated Plan on AI.[67] This 2021 Coordinated Plan is the next step for the EU in creating global leadership in trustworthy AI. It builds on the strong collaboration between the EC and Member States established during the 2018 Coordinated Plan. The report confirms that NLP is one of the most rapidly advancing fields of AI. The combination of the first-ever legal framework on AI[68] and the new coordinated plan will guarantee the safety and fundamental rights of people and businesses, while strengthening AI uptake, investment and innovation across the EU.

Although AI is often considered a general-purpose technology, LT is highlighted for having systems with the ability to analyze, understand and generate information expressed in natural language. These capabilities are crucial for improving human-computer interaction. In fact, different reports from AI Watch[69] place LT as one of the most important areas within AI. *AI Watch – Defining Artificial Intelligence*[70] proposes a taxonomy with a comprehensive collection of areas that represents AI from three target perspectives: policy, research and industry. NLP is classified as the main task of the communication domain. In fact, NLP is considered a subdomain of AI in several national strategies, encompassing applications such as text generation, text mining, text classification, machine translation and speech recognition.

For instance, *AI Watch – Artificial Intelligence in public services*[71] presents the landscape of AI use in public services, identifying 230 cases from which to extract emerging trends and examples of current AI usage. The report suggests ten AI typologies aligned with the taxon-

[64] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[65] https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data
[66] 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/
[67] https://digital-strategy.ec.europa.eu/en/library/new-coordinated-plan-artificial-intelligence
[68] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence
[69] https://knowledge4policy.ec.europa.eu/ai-watch_en
[70] https://ec.europa.eu/jrc/en/publication/ai-watch-defining-artificial-intelligence
[71] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/ai-watch-artificial-intelligence-public-services

omy proposed by AI Watch, focusing on ML and other approaches. The authors highlight that limitations of the study include *translation issues, for the documents not available in the English language*. The observed 230 cases show a certain variety: the majority (51 cases) in "Chatbots, Intelligent Digital Assistants, Virtual Agents and Recommendation Systems"; 36 cases of applications in "Predictive Analytics, Pattern Recognition, Simulation and Data Visualisation" and 17 cases of "Machine Learning and Deep Learning". The class "Natural Language Processing, Text Mining and Speech Analysis" encompasses 19 cases. That is, well over half of the total of 230 are very closely related to LT.

According to a recent study from Eurostat,[72] in 2020, 7% of enterprises in the EU with at least ten employees used AI applications. While 2% of the enterprises used ML to analyse big data internally, 1% analysed big data internally with the help of LT. A chat service, where a chatbot or virtual agent generated natural language replies to customers, was used in 2% of the enterprises. The same proportion of enterprises, 2%, used service robots, for example to carry out cleaning, dangerous or repetitive tasks such as removing poisonous substances, sorting items in the warehouse, helping customers in shopping or at payment points etc.[73]

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 to better understand small and medium-sized enterprises' (SMEs) interests, needs, existing solutions and use cases for website translation.[74] The survey on multilingual websites aimed to further analyse the language barriers across the EU Member States. Over 1,000 SMEs replied to the consultation, 75% of which expressed interest in participating in the EC's subsequent pilot testing to make their website automatically multilingual. Overall, the consultation identified specific market needs that could be addressed by European language service providers complemented by public solutions, such as those based on eTranslation.

A recently released Eurobarometer survey on "European citizens' knowledge and attitudes towards science and technology" shows that 9 in 10 EU citizens (86%) believe that the overall influence of science and technology is positive.[75] EU citizens expect a range of technologies currently under development to have a positive effect on their way of life in the next 20 years: notably, solar energy (92%), vaccines and combating infectious diseases (86%) and AI (61%). Respondents most often mention health and medical care and the fight against climate change when asked in which areas research and innovation can make a difference.

In summary, all main AI related reports and policy initiatives highlight the extensive impact of AI in society. In these reports, the relevance of LT (together with Vision and Robotics) is assessed as one of the most important functional applications of AI.

4. Language Technology in European Initiatives

The European LT community, as currently represented in and by the EU projects ELE (European Language Equality) and ELG (European Language Grid) is committed to doing the research, development and deployment of ground-breaking and novel technologies in order to successfully turn a linguistically fragmented Europe into a truly unified and inclusive one, fully supporting the rich and diverse linguistic cultural heritage from broadly spoken languages to minority and regional languages as well as the languages of immigrants and important trade partners, benefiting the European citizen, European industry and European society. Multilingualism is within the scope of a series of EU policy areas, including culture, education, the economy, the Digital Single Market (DSM), lifelong learning, employment, social

[72] https://ec.europa.eu/eurostat/web/main/home
[73] https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20210413-1
[74] https://digital-strategy.ec.europa.eu/en/library/report-sme-survey-multilingual-websites
[75] https://europa.eu/eurobarometer/surveys/detail/2237

inclusion, competitiveness, youth, civil society, mobility, research and media. More attention needs to be paid to removing barriers to intercultural and inter-linguistic dialogue, and to stimulating mutual understanding. According to the recent European Parliament (2018) resolution (point D),

> 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.

The first two pages of the *Language equality in the digital age* resolution (European Parliament, 2018) are shown in Figure 3.

The EU has already dedicated efforts and funding to advancing LT, in particular through Horizon 2020 and Connecting Europe Facility (CEF).[76] The MT service etranslation is a key achievement of the CEF programme.[77]

European Parliament
2014-2019

TEXTS ADOPTED

P8_TA(2018)0332

Language equality in the digital age

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))

The European Parliament,

– having regard to Articles 2 and 3(3) of the Treaty on the Functioning of the European Union (TFEU),

– having regard to Articles 21(1) and 22 of the Charter of Fundamental Rights of the European Union,

– having regard to the 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage,

– having regard to Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information[1],

– having regard to Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information[2],

– having regard to Decision (EU) 2015/2240 of the European Parliament and of the Council of 25 November 2015 establishing a programme on interoperability solutions and common frameworks for European public administrations, businesses and citizens (ISA2 programme) as a means for modernising the public sector[3],

– having regard to the Council resolution of 21 November 2008 on a European strategy for multilingualism (2008/C 320/01)[4],

– having regard to the Council decision of 3 December 2013 establishing the specific programme implementing Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) and repealing Decisions 2006/971/EC, 2006/972/EC, 2006/973/EC, 2006/974/EC and 2006/975/EC[1],

– having regard to the UN Convention on the Rights of Persons with Disabilities (UN CRPD), ratified by the EU in 2010,

– having regard to the Commission communication of 18 September 2008 entitled 'Multilingualism: an asset for Europe and a shared commitment' (COM(2008)0566),

– having regard to the Commission communication of 26 August 2010 entitled 'A Digital Agenda for Europe' (COM(2010)0245),

– having regard to the Commission communication of 11 January 2012 entitled 'A coherent framework for building trust in the Digital Single Market for e-commerce' (COM(2011)0942),

– having regard to the Commission communication of 6 May 2015 entitled 'A Digital Single Market Strategy for Europe' (COM(2015)0192),

– having regard to the opinion of the European Economic and Social Committee on the communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 'A Digital Agenda for Europe' (COM(2010)0245)[2],

– having regard to the Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace adopted by the UNESCO General Conference at its 32nd session in Paris on 15 October 2003,

– having regard to the Special Eurobarometer 386 report entitled 'Europeans and their Languages', published in June 2012,

– having regard to the Presidency conclusions of the Barcelona European Council of 15 and 16 March 2002 (SN 100/1/02 REV 1),

– having regard to its resolution of 17 June 1988 on sign languages for the deaf[3],

– having regard to its resolution of 14 January 2004 on Preserving and promoting cultural diversity: the role of the European regions and international organisations such as UNESCO and the Council of Europe[4], and to its resolution of 4 September 2003 on European regional and lesser-used languages – the languages of minorities in the EU – in the context of enlargement and cultural diversity[5],

– having regard to its resolution of 24 March 2009 on Multilingualism: an asset for Europe and a shared commitment[6],

1 OJ L 345, 31.12.2003, p. 90.
2 OJ L 175, 27.6.2013, p. 1.
3 OJ L 318, 4.12.2015, p. 1.
4 OJ C 320, 16.12.2008, p. 1.

1 OJ L 347, 20.12.2013, p. 965.
2 OJ C 54, 19.2.2011, p. 58.
3 OJ C 187, 18.7.1988, p. 236.
4 OJ C 92 E, 16.4.2004, p. 322.
5 OJ C 76 E, 25.3.2004, p. 374.
6 OJ C 117 E, 6.5.2010, p. 59.

Figure 3: The first two pages of the 2018 EP Resolution *Language equality in the digital age*.

While the official EU languages are granted equal status politically, technologically they are far from equally supported. In this respect, the following studies should be highlighted as they represent key contributions to the debate in Europe on these issues:

- *The FLaReNet Strategic Language Resource Agenda* (Soria et al., 2014) – the Strategic Language Resource Agenda by the European LT community. Among other important recommendations for developing, documenting, sharing and reusing language resources,

76 https://ec.europa.eu/digital-single-market/en/language-technologies
77 https://ec.europa.eu/education/knowledge-centre-interpretation/eu-initiatives-language-technologies_en

it is stressed the need of defining and developing Basic Language Resource Kits[78] (BLARK) for all languages.

- *META-NET White Paper Series: Europe's Languages in the Digital Age* (Rehm and Uszkoreit, 2012; Rehm et al., 2014c) – the first systematic study about the technology support of Europe's languages.
- *META-NET Strategic Research Agenda for Multilingual Europe 2020* (Rehm and Uszkoreit, 2013; Rehm et al., 2014a,b) – the first SRA by the European LT community.
- *Language Equality in the digital age – Towards a Human Language Project*, commissioned by the Science and Technology Options Assessment (STOA) of the European Parliament (STOA, 2017).
- The Strategic Research and Innovation Agenda *Language Technologies for Multilingual Europe – Towards a Human Language Project* (Rehm, 2017) published by the European Language Technology community.
- The STOA report (mentioned above), other strategic documents and additional input was used to prepare the 2018 EP resolution *Language equality in the digital age* (European Parliament, 2018).
- The Rehm and Hegele (2018) survey representing the voices of more than 600 respondents from more than 50 countries working on LT.
- Rehm et al. (2020b) emphasise the necessity of a programme tailored specifically to Europe's needs and demands.

The STOA study presented in a joint ITRE and CULT committee meeting and to the STOA panel, led to the preparation of a joint Resolution of the European Parliament (European Parliament, 2018). In a plenary meeting on 11 September 2018, the European Parliament adopted this joint ITRE/CULT resolution, *Language equality in the digital age*[79] (European Parliament, 2018) with an overwhelming majority of 592 votes in favour, 45 against and 44 abstentions. The adoption of the resolution with such a landslide majority demonstrates the importance and relevance of the topic. The ten-page resolution includes more than 40 recommendations, structured into the following 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". It includes many recommendations that are highly relevant for the topic of digital language equality. The resolution (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)

[78] http://www.blark.org
[79] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html

- "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 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)

The ELE project addresses the EP recommendations and lays the foundations for a systematic plan and roadmap for making digital language equality a reality in Europe by 2030.[80]

In the final statement of the hearing on the the EP Resolution (European Parliament, 2018) on 11 September 2018, EC Commissioner Corina Crețu acknowledged the importance and relevance of the resolution, saying

> 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.

This is in line with previous public appeals made in 2016 by former European Commission Vice President Andrus Ansip, *How multilingual is Europe's Digital Single Market?*[81] and in 2017 by Director General Roberto Viola (DG Connect) *Multilingualism in the Digital Age: a barrier or an opportunity*[82] for the need to strengthen multilingualism through technologies. Current EC initiatives, such as ELG (Rehm et al., 2020a, 2021) and the eTranslation building

[80] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/langeq-2020
[81] https://ec.europa.eu/commission/commissioners/2014-2019/ansip/blog/how-multilingual-europes-digital-single-market_en
[82] https://ec.europa.eu/digital-single-market/en/blog/multilingualism-digital-age-barrier-or-opportunity

block of CEF, as well as ongoing investment in MT, do contribute to continuous progress. LT *for* Europe *made in* Europe is the key. Not only will it strengthen Europe's place in the pole position of research excellence, but it will contribute to future European cross-border and cross-language communication, economic growth and social stability.

The *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem* from 2019 provides an analysis of the LT market of the EU (including Norway and Iceland) and the adoption of LT by public administrations, both EU-level and national.[83] As for weaknesses and threats, this report considers that the EU industry is fragmented with many small players struggling to find a place in the market in order to compete with the global players, which dominate the market and upon which European businesses and public sector have become dependent. While the research position of the EU is weakening, the global US players have a large competitive advantage in terms of research capacities, computing resources and available data. Fortunately, the European High Performance Computing Joint Undertaking[84] (EuroHPC JU), a joint initiative between the EU, European countries and private partners, is developing a World Class Supercomputing Ecosystem in Europe. In fact, the first call for proposals for EuroHPC JU regular access mode has recently open.[85] In addition, the first calls for this programme include a call for a Language Data Space.[86]

They distort the market, for instance by providing free MT services, even though MT is not their core business. They also have larger amounts of data at their disposal, because of copyright disparities between the EU (requirement of explicit permission by European entities) and the US (fair use copyright exception), and because the intensive use of their popular systems allows them to collect a lot of user data. In contrast, the EU industry is experienced with small and complex languages, serving limited markets and restricted business opportunities, and the amount of accessible data for these languages is low. As for strengths and opportunities, European MT developers have been successful in deploying services for the public sector through the support of EU-funded programmes. In the market, three deficiencies can be observed, providing opportunities for the EU: 1) There are gaps in the offering for small and complex languages; 2) There is a lack of domain-specific and application-specific MT; 3) US global players pay little attention to low-resource languages as well as to security and privacy issues. In fact, as stated in the 2019 document *My Europe. My language. With language technologies made in the EU*,[87] LT offers opportunities to reduce language barriers across Europe and in the DSM at the intersection of Big Data, AI and HPC.

The European Language Resource Coordination (ELRC) White Paper *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters* identified that the main challenge by far is the lack of appreciation of the value of language data.[88] The recommendations for the European and national policy level include:

- Updating the Open Data Directive (2019/1024/EU) that should reference language data as a high-value data category.[89]
- Conducting of a study on the value of language data to identify and quantify the value of language data for citizens, public administrations and businesses.

[83] https://op.europa.eu/en/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en/format-PDF/source-106906783
[84] https://eurohpc-ju.europa.eu
[85] https://eurohpc-ju.europa.eu/news/access-eurohpc-supercomputers-now-open
[86] https://digital-strategy.ec.europa.eu/en/activities/work-programmes-digital
[87] https://ec.europa.eu/digital-single-market/en/news/my-europe-my-language-language-technologies-made-eu-brochure
[88] https://lr-coordination.eu/sites/default/files/Documents/ELRCWhitePaper.pdf
[89] https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

- Updating national policies (e. g., national Open Data policy, digital agenda or strategy 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 assessing translation practices in public administrations on all administrative levels.

Along with the importance of language equality in the digital age, another aspect to take into account is the accessibility of information in a multimodal context, not only with regard to its format, but also the understanding of its content, through the simplification, summarisation or adaptation of concepts, sentences, paragraphs or texts, contributing to the development of inclusive digital societies. For this task, LT is essential, as reflected in the *Report on the Joint Stakeholder Consultation on Research and Innovation in Web Accessibility and Language Technologies*,[90] in which one of the actions to be carried out is related to the development of systems capable of adapting and personalizing digital content according to the needs of each person, in terms of accessibility and language. For instance, research efforts on sign languages are being addressed but must be significantly extended since more and more sign languages are becoming recognised as official national languages.

In parallel, the Coordinated Artificial Intelligence Plan proposed by the European Commission, for the period 2021-2027, aims to move Europe into the pole position when it comes to developing, using and exploiting AI technologies.[91] This plan has a minimum annual investment by the EU of one billion Euros in Horizon Europe and Digital Europe, but whose most ambitious objective is to reach 20 billion Euros a year between public and private investments. This plan focuses on four key areas: 1) increasing investment in AI; 2) the availability of data; 3) the promotion of talent; and 4) ensuring security, ethics and trust in AI. Following this plan, the EC urged its Member States to develop and coordinate their own national AI strategies, the analysis and comparison of which is included in the report *AI Watch – National strategies on Artificial Intelligence: A European perspective in 2019* (see Section 5).[92]

The EC has recently established a public-private partnership (PPP) in the area of AI.[93] In September 2020, the Data, AI and Robotics Partnership including BDVA,[94] euRobotics,[95] ELLIS,[96] CLAIRE,[97] and EurAI[98] presented a third version of their SRIDA (Strategic Research, Innovation and Deployment Agenda). AI, Data and Robotics are transversal technologies and cut across sectors affecting many actors in the value chain. There is widespread acceptance that AI, Data and Robotics will have significant impact on all economic sectors and on the United Nations' sustainable development goals.[99] According to this SRIDA, "NLP has particular resonance within Europe's multi-lingual landscape and offers the potential to harmonise human interaction." The document also agrees that the "most well-known approaches to-date in machine learning, that are responsible for the recent successes of AI are based on deep neural networks. It has resulted in a world-wide dissemination in computer vision,

[90] https://ec.europa.eu/digital-single-market/en/news/report-joint-stakeholder-consultation-research-and-innovation-web-accessibility-and-language-0
[91] https://knowledge4policy.ec.europa.eu/ai-watch/coordinated-action-plan-ai_en
[92] https://ec.europa.eu/jrc/en/publication/ai-watch-national-strategies-artificial-intelligence-european-perspective-2019
[93] https://adr-association.eu
[94] https://www.bdva.eu
[95] https://www.eu-robotics.net
[96] https://ellis.eu
[97] https://claire-ai.org
[98] https://eurai.org
[99] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

signal processing, NLP, where new application domains appear continuously." However, although the PPP is including 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. Curry et al. (2021) summarize the collective effort undertaken by the European data community as part of the Big Data Value PPP between the EC and the BDVA to establish the Technical Research Priorities for Big Data. According to this report:[100]

> 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 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.

European Learning and Intelligence Systems Excellence[101] (ELISE) has recently published its SRA, setting out a roadmap for European leadership in the development of safe and effective AI systems. ELISE is a consortium of AI research hubs that connects Europe's leading researchers in ML and AI. The report agrees that progress in ML has also enabled further advances in fields such as NLP and computer vision, which contribute to the creation of AI systems. Among other goals, they also aim to build systems for general-purpose NLU and generation.

The New European Media[102] (NEM) Strategic Research and Innovation Agenda[103] also highlights the importance of LT as an innovation content enabler:

> In the future AI based tools will be developed, such as automatic translation from speech to subtitles, from text to Sign Language, and from Sign Language to text. These actions are essential to maintain Europe's position as the World leader in accessibility and for social and societal challenges.

Two other European initiatives[104] in which LT is present to a greater or lesser extent deserve to be mentioned here as part of this report:[105] 1) the Horizon 2020 Programme[106] (already completed), in which LT is embedded within research and innovation in the field of information technologies, content technologies, multilingual internet and AI, and 2) the Connecting Europe Facility,[107] through which MT tools (eTranslation[108]) or tools for the management of thesauri and glossaries have been developed (VocBench[109]).

More recently, the EP's CULT Committee adopted a resolution on AI in the cultural, creative and educational sector[110] in which multilingual and linguistic diversity is also taken

[100] https://elements-of-big-data-value.eu/research-priorities-for-big-data-value/#page-content
[101] https://www.elise-ai.eu
[102] https://nem-initiative.org
[103] https://nem-initiative.org/wp-content/uploads/2020/06/nem-strategic-research-and-innovation-agenda-2020.pdf?x98588
[104] https://ec.europa.eu/education/knowledge-centre-interpretation/eu-initiatives-language-technologies_es
[105] https://ec.europa.eu/digital-single-market/en/language-technologies
[106] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/information-and-communication-technologies
[107] https://ec.europa.eu/digital-single-market/en/connecting-europe-facility
[108] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[109] https://ec.europa.eu/isa2/solutions/vocbench3_en
[110] https://www.europarl.europa.eu/news/en/press-room/20210311IPR99709/ai-technologies-must-prevent-discrimination-and-protect-diversity

into account. Regarding linguistic diversity, the resolution calls for: 1) AI technologies to be regulated and trained in order to protect non-discrimination, gender equality, pluralism, as well as cultural and linguistic diversity; 2) specific indicators to measure diversity and ensure that European works are being promoted in order to prevent recommendations that negatively affect the EU's cultural and linguistic diversity, and 3) the establishment of an ethical framework on the use of AI technologies in EU media to ensure people have access to culturally and linguistically diverse content. Such a framework should also address the misuse of AI to disseminate fake news and disinformation.[111] According to rapporteur Sabine Verheyen (MEP, Germany):

> Developing quality and inclusive data systems for use in deep learning is vital, as is a clear ethical framework to ensure access to culturally and linguistically diverse content.

The CULT resolution will be presented soon to the plenary of the European Parliament for a vote. 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 is scheduled to be published in early 2022.

In summary, Europe's multilingual nature is also one of the main obstacles to a truly connected, cross-lingual communication and information space. While language diversity is at the core of Europe identity and multilingual society, 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).[112] 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 been facing increased competition from other continents, especially with regard to recent breakthroughs in AI. These scientific breakthroughs 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 research abroad. Europe should invest in retaining and attracting young high potential researchers. Our continent is in need of powerful LT *made in* Europe *for* Europe and *for* all European citizens, tailored to our specific cultures and societal as well as economic demands in order to successfully turn a linguistically fragmented Europe into a truly unified and inclusive one. This ambitious but worthy effort involves supporting the rich and diverse linguistic cultural heritage from broadly spoken languages to minority and regional languages as well as the languages of immigrants and important trade partners, benefiting European citizens, European industry and European society.

5. National Language Technology Initiatives in Europe

Rehm et al. (2020b) present an overview of various European LT and AI reports. As part of the European Language Grid[113] (Rehm et al., 2020a, 2021), the 32 ELG National Competence Centres contributed information about the national funding situation for AI- and LT-related topics in their countries (see Rehm et al., 2014a,b, for a previous overview analysis).

Similarly, in 2018, as part of the Spanish *Plan for the Advancement of Language Technology*, an analysis of the LT landscape in Europe was presented.[114] European strategies, policies and

[111] https://op.europa.eu/en/publication-detail/-/publication/b8722bec-81be-11e9-9f05-01aa75ed71a1
[112] http://www.unesco.org/languages-atlas/index.php?hl=en&page=atlasmap
[113] https://www.european-language-grid.eu
[114] https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/tecnologias-del-lenguaje-en-Europa.aspx

programmes to support LT were identified. At the national level, they concluded that there are or have been only very few dedicated national programmes designed to finance projects related to LT. In contrast, in most countries funding for the development of LT is provided through generic R&D&I calls. At the European level, so far LT has received better support through calls in different programmes: FP7, H2020, CEF Telecom, CIP ICT-PSP, EUREKA and EUROSTARS, among others. However, the authors note that in the most recent programmes funding for LT projects has been gradually reduced.

If we compare both studies, we can observe a small increase in the number of language-centric AI initiatives in the last couple of years. In the following, we summarise the main results presented by Rehm et al. (2020b) and we update the summary table (see Table 1) with the most recent reports and initiatives: AI for Belgium, the draft Concept for the Development of AI in Bulgaria, the National AI strategy of Cyprus, the Digital Transformation Strategy 2020-2025 of Greece, the Icelandic AI strategy, *AI - Here for Good: National Artificial Intelligence Strategy for Ireland*, Latvia's AI reports, AI: a strategic vision for Luxembourg, the National Strategy for AI in Norway, Romania in the era of AI and Slovakia's strategy of the Digital Transformation 2030. The effort of all European countries to be in line with the EC, which considers AI an area of strategic importance is noteworthy. In December 2018, the EC and the Member States published the "Coordinated Plan on Artificial Intelligence", COM(2018)795, on the development of AI in the EU. The number of EU countries with an AI strategy (29 out of 30, 97%) demonstrates the success of the plan. Only Croatia has no official AI strategy as of yet.

The *AI Watch National strategies on AI: A European perspective in 2019* report also analyses the EU national AI strategies to identify areas for synergies and collaboration. It identifies several policy areas: human capital, from lab to market, networking, infrastructure, regulation. LT is mentioned as part of the Danish, Latvian, Maltese, Portuguese, Slovakian, Spanish and Swedish initiatives, so, as Rehm et al. (2020b) mention, LT is finally generating some momentum. LT is considered key, and language understanding is seen as one of the next generations of innovative AI technologies (STOA, 2017). For that, it is indispensable to set aside funding exclusively for LT. According to Rehm et al. (2020b), only four of the 30 surveyed countries do not have some type of LT funding. Four countries have programmes dedicated to LT (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. The 2021 edition of the AI Watch National Strategies on AI: A European perspective report not only highlights various priority sectors (manufacturing, agriculture, healthcare, transport and energy) with a high potential for AI applications, but also identifies a number of countries prioritising LT in their AI strategies as key to introduce LT-based AI applications in public services such as interactive dialogue systems and personal virtual assistants.[115]

In summary, there are or have been only very few dedicated nationally financed programmes related to LT. In contrast, national funding for the development of LT is provided through generic R&D&I calls, not specific programmes. Moreover, only 12 European countries out of the 30 studied explicitly consider LT within their national policy initiatives.

6. Non-EU National Initiatives

The *AI Index Report 2021*[116] includes a chapter on AI Policy and National Strategies with pointers to available national initiatives organized by year. The chapter highlights:

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

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

	LT-related funding			Artificial Intelligence	
	None at all	Some funding	Dedicated LT programme	AI strategy	LT funding through AI
Austria	X			X	
Belgium		X		X	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		X	
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		X	
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.

- 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.
- The launch of the Global Partnership on AI (GPAI) and OECD AI Policy Observatory and Network of Experts on AI in 2020 promoted intergovernmental efforts to work together to support the development of AI for all.
- In the US, the 116th Congress was the most AI-focused congressional session in history. The number of mentions of AI by this Congress in legislation, committee reports, and Congressional Research Service (CRS) reports was more than triple that of the 115th Congress.

Among all the published strategies, we would like to mention some of the non-EU ones, namely the AI strategies from China, India, Mexico, United Kingdom and the United States.

In 2017, China, along with Canada, Japan and others, published their national strategy for AI. According to the AI index report (Zhang et al., 2021),

> China's AI strategy is one of the most comprehensive in the world. It encompasses areas including R&D&I and talent development through education and skills acquisition, as well as ethical norms and implications for national security. It sets specific targets, including bringing the AI industry in line with competitors by 2020; becoming the global leader in fields such as unmanned aerial vehicles, voice and image recognition, and others by 2025; and emerging as the primary center for AI innovation by 2030.

What is especially remarkable is the presence of LT in the AI strategy of China. As part of the objectives designed to significantly enhance China's cadre of AI talent and its university AI curricula by 2030, NLU is considered as one of the key technologies to study as well as to promote the innovation. The creation of more AI-oriented teaching materials and open online curricula in NLP is also considered one of the key tasks.

In 2018, India, Mexico and the UK published their AI strategies. The UK strategy emphasizes a strong partnership between business, academia, and government. It identifies five foundations for the industrial strategy: to become the world's most innovative economy; to offer good jobs and greater earning power for all; to upgrade the UK's infrastructure; to offer a business environment to grow in business; and to build prosperous communities across the UK. As such, they consider it indispensable to invest in increasing the size of the AI workforce. As a result, they run a pilot programme for under-18 year olds across the UK, to encourage them to consider a career in the AI sector, in which NLP is explicitly mentioned.

Mexico announced the first strategy in Latin America (Del Pozo et al., 2020) and is focused on developing a strong governance framework, mapping the needs of AI in various industries, and identifying governmental best practices. The *Towards a strategy for AI focused on Language Technologies in Spain* report also analysed the situation in Mexico and, as regards NLP, highlights the number of indigenous languages as well as the lack of NLP applied to them. The report stresses that the situation is similar throughout Latin America.

Regarding India, the AI Index report highlights the focus on economic growth and ways to leverage AI to increase social inclusion, while also promoting research to address important issues such as ethics, bias, and privacy related to AI. As an example of AI solutions to account for cultural nuances and diversity, the *AI in India: A Policy Agenda* report mentions natural language voice recognition to account for the diversity in languages and digital skills in the Indian context. Similarly, the national strategy on AI considers the multilingual reality of the country a crucial aspect to achieve technology leadership in AI. As an example, they mention the development of an advanced NLP infrastructure for the languages of India. They recommend the creation of annotated data sets for their languages to add incremental value to existing services ranging from e-commerce to agricultural advisory.

Finally, in 2019, the US presented the *American AI Initiative*. It prioritizes the need for the federal government to invest in AI R&D&I, reduce barriers to federal resources, and ensures technical standards for the safe development, testing, and deployment of AI technologies. The White House emphasizes developing an AI-ready workforce and signals a commitment to collaborating with foreign partners while promoting US leadership in AI. The National Security Commission on AI report[117] also highlights many times the crucial role of LT. In this report, NLU appears as one of the six *Uses for Deployed AI Today*.

In summary, all non-EU national AI initiatives explicitly consider LT and NLP as one of the most relevant strategic areas on which to focus.

7. SWOT Analysis

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

7.1. 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 machine translation are available.
- Existence of multiple national and European LT research networks, associations, communities and other relevant stakeholders whose objective is to promote all kinds of activities related to research, development, education and industry in the field of LT, both nationally and internationally.
- Existence of unique, valuable and potentially very useful data resources that can be exploited by current LT. An enormous amount of 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).

7.2. 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.

[117] https://www.nscai.gov/2021-final-report

- 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 for 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 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.

7.3. 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 neighboring countries through good partnerships based on the use of LT customized to other languages.
- Growing trend for the LT market and industry in Europe regarding the exploitation of digital resources and data of linguistic interest. Digitisation is one of the key means to generate new economic growth.
- Consolidation of a competitive LT industry that harnesses the potential of research and academia both in educating well-trained LT professionals and in transferring research results to industry and public administrations.
- Increasing interest in higher education to 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 of 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.

7.4. 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.

8. Recommendations

In line with the final recommendations of the EP *Language equality in the digital age* resolution (European Parliament, 2018), the main recommendations put forward by the reports, documents and initiatives analysed in this deliverable can be summarized as follows.

- To reinforce European leadership in LT by creating a specific programme tailored to Europe's needs and demands, i. e., to establish a large-scale, long-term coordinated funding programme for research, development, innovation and education in the field of LT, at European, national and regional levels, tailored specifically to Europe's needs and demands to tackle *deep NLU*. This will increase efficiency by sharing knowledge, infrastructures and resources, with a view to developing innovative technologies and services, in order to achieve the next scientific breakthrough in this area and help reduce the technology gap between European languages with the participation of research centres, academic experts, enterprises and other relevant stakeholders. Crucially, such a long-term programme must involve significantly improved coordination between European LT research and industry.
- To safeguard sufficient funding to support the new technological approaches, based on increased computational power and better access to sizeable amounts of data, in order to foster the development of deep-learning neural networks and other emerging technologies which make LT a real solution to the problem of language barriers.
- To create 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.

- Within the EC, to allocate the area of *multilingualism and language technology* to the portfolio of a Commissioner. This Commissioner should be tasked with promoting linguistic diversity and equality at EU level, given the importance of linguistic diversity and the opportunity of digital language equality for the future of Europe.
- To ensure comprehensive EU-level legal protection for the more than 60 regional and minority languages, recognition of the collective rights of national and linguistic minorities in the digital world (including sign languages), and mother-tongue teaching for speakers of official and non-official languages of the EU.
- To ensure the development of LT for official EU languages which are less widely spoken and the launch of special EU-level actions (policy, funding, research, education) to include and promote regional and minority languages in such development.
- To coordinate the development of comprehensive language-related policies by the Member States and to allocate funding and resources to promote and facilitate linguistic diversity and multilingualism in the digital sphere, including languages that are less widely spoken.
- To develop strategies and policy actions by the EC and Member States to facilitate multilingualism in the digital market; to define the minimum language resources and capacities that all European languages should possess, defining and developing a BLARK-like minimum set of LT resources.
- To include language data as a high-value data category in the Open Data Directive (2019/1024/EU). To develop common policy actions and clear protocols for language data sharing by the public administration at all levels.
- To create a centre for linguistic diversity that will strengthen awareness of the importance of lesser-used, regional and minority languages, including in the sphere of language technologies.
- To reduce the technology gap between languages by strengthening knowledge and technology transfer.
- To strengthen and reinforce the European Language Grid as the primary European LT platform, with representatives from all European languages, so as to enable and foster the sharing of language technology-related resources, services and open source code packages between stakeholders from research and industry, while ensuring that any funding scheme can both work with and be accessed by the open-source community.
- To periodically update the META-NET White Paper series on the status of technology support for Europe's languages, on resources for all European languages, on information about language barriers and on policies related to the topic, with a view to enabling the assessment and development of LT policies.
- To aim European education policies at retaining talent in Europe, analyse the current educational needs related to LT and, based on this, to provide guidelines for the implementation of cohesive joint action at European level, and to raise awareness among schoolchildren and students of the career opportunities in the LT industry, including the language-centric AI industry.
- To promote increased participation of women and representatives of ethnic and social minorities in the field of European studies on LT, as a decisive factor in the development of fair and inclusive research and innovation.

- To develop actions and appropriate funding with the aim of enabling and empowering European SMEs and startups to easily access and use LTs in order to grow their businesses online by accessing new markets and development opportunities, thereby boosting their levels of innovation and creating jobs.
- To secure a European leadership position in the field of language-centric AI since EU companies are the best placed to provide solutions tailored to our specific cultural, societal and economic needs.
- To create the necessary appealing conditions to attract and retain international LT experts in EU flagship research centers and institutes in universities and companies.
- To ensure trust in LT thanks to a regulatory framework for the development, implementation and deployment of reliable LT technologies, in line with established legal and ethical AI principles and good practices within the EU and Member States.
- To engage administrations at all levels, and to use existing free and open-source LT, in order to improve the accessibility of those services.

As the ELE project progresses, these preliminary recommendations will be revisited, revised and extended on a regular basis, culminating in our final recommendations in the form of the strategic agenda and roadmap.

Acknowledgements

The authors would like to thank the whole ELE consortium for their valuable contributions, especially regarding the input, comments and suggestions for the national and local LT and AI initiatives and reports.

References

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.

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.

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.

Claudia May Del Pozo, Constanza Gómez Mont, and Cristina Martínez Pinto. *Artificial Intelligence in Mexico: A National Agenda. The Agenda in Brief.* Mexico: IA2030Mx, 2020.

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.

ESFRI-Roadmap. Strategy report on research infrastructures - roadmap 2018. *EU, roadmap2018. esfri. eu*, 2018.

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.

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.

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

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 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 Odijk, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Michael Rosner, Bolette 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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1517–1524, Reykjavik, Iceland, 2014a. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/405_Paper.pdf.

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 Odijk, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Michael Rosner, Bolette 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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1517–1524, Reykjavik, Iceland, 2014b. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/405_Paper.pdf.

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, 2014c.

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 Hajič, 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 *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France, 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.413.

Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajič, 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 Skadiņa, 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 *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3322–3332, Marseille, France, 2020b. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.407.

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, Jose Manuel Gomez-Perez, 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 Hamrlova, Dusan Varis, Lukas Kacena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Julija Melnika, 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*, pages 221–230, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-demos.26.

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 Computational 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, 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 (original version).

	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 (November 2021).