EUROPEAN
LANGUAGE
EQUALITY

D1.11

Report on the English Language

Authors	Diana Maynard, Joanna Wright, Mark A. Greenwood, Kalina Bontcheva
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.11
Deliverable title	Report on the English Language
Type	Report
Number of pages	21
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Diana Maynard, Joanna Wright, Mark A. Greenwood, Kalina Bontcheva
Reviewers	John Judge, Ruben Urizar
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2022 ELE Consortium

Consortium

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

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CLTL	Cross-Lingual Transfer Learning
DLE	Digital Language Equality
EU	European Union
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
EPSRC	Engineering and Physical Sciences Research Council
EU	Europan Union
GPU	Graphics Processing Unit
HPC	High-Performance Computing
LDC	Linguistic Data Consortium
LR	Language Resources/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NCC	National Competence Centre
NLG	Natural Language Generation
NLP	Natural Language Processing
PoS	Part-of-Speech
SFI	Science Foundation Ireland
SR	Speaker Recognition
TTS	Text-to-speech

Abstract

This report is part of a series describing the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and aims to delineate the current state of affairs for each of the European languages covered. Additionally – and most importantly – the series aims to identify the gaps as well as the factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based proposal of required measures for achieving Digital Language Equality in Europe by 2030.

This report focuses on the status of the English language, primarily acting as a benchmark for the level of technological support that other European languages could receive. While it is extremely unlikely that any other European language will reach this level, due to the continuing development of support for English, and thus serves as a moving goalpost, nevertheless it provides a good criterion for relative assessment. English is a truly international language, with the highest number of speakers in the world for any language, and serving as the primary language of international discourse, as well as the lingua franca in many professional contexts. It is almost impossible to measure accurately the numbers of resources and language tools for English, since these are spread so widely, but this report makes some estimates and gives some benchmark figures for comparison with other languages about those resources available in the largest and most well-known repositories, such as the ELG, ELRA and LDC collections.

This report first provides some background information, before a brief description of English as a language, including its geography, history, typology, and use in the digital world. It then discusses the current state of language technology for English, looking at both various kinds of language and speech processing tools, and language resources such as corpora, dictionaries and grammars. It also discusses the role of projects, initiatives and national funding initiatives. Finally, a cross-language comparison is made, before some conclusions are drawn.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but additionally – and most importantly – to identify the gaps as well as the factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners, who are experts in more than 30 European languages, have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed in the frame of the European Language Equality (ELE) project.[2] With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

[2] https://european-language-equality.eu

ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

2 The English Language in the Digital Age

2.1 General Facts

English is a truly international language, due in no small part to the worldwide influence of the British Empire since the 17th century, and later to the influence of the United States. It has become the primary language of international discourse and is the lingua franca in many professional contexts, as well as in a number of geographic regions with diverse native languages.

According to the 2021 edition of Ethnologue,[3] it is the most spoken language in the world (in terms of number of speakers), with an estimated 1.348 billion total speakers (closely followed by Mandarin Chinese with a total of 1.120 billion speakers). For comparison, the language with the third highest total number of speakers is Hindi with 600 million speakers.

English is also the most widely spoken second language in the world according to Ethnologue. There are more than twice as many people who speak English as a second language than there are native speakers, with a total of 369.9 million first language (L1) speakers and 978.2 million second language (L2) speakers. In comparison, Mandarin Chinese has the highest number of native speakers (921 million, more than double that of English), while Spanish has the second highest number (471 million). Hindi has 342 million native speakers.

English is the majority native language in a group of five countries often known as the (core) Anglosphere, comprising the United Kingdom, the United States, Canada, Australia, New Zealand. It is also the primary native language in Ireland, an official (and the primary) language of Singapore, and widely spoken in some areas of the Caribbean, Africa, South Asia, Southeast Asia and Oceania, as depicted in Figure 1.[4] It is a co-official language of the United Nations, the European Union and many other world and regional international organisations. Of the Germanic languages,it is the most widely spoken, accounting for at least 70% of speakers.

English-speaking countries are often defined according to a (constantly evolving) three-circles model (Svartvik and Leech, 2006) based on historical language evolution. The “inner circle” countries have large communities of native speakers of English, and include the Anglosphere countries as well as Ireland and South Africa. The “outer circle” countries have only small communities of native English speakers, but typically use English as a second language in education or broadcasting or for local official purposes, such as India and Nigeria. Finally, “expanding circle” countries are those where English is frequently learned as a foreign language, such as China and many European countries.

Finally, English is classed as a pluricentric language (Clyne, 1992), meaning that it has no single standard codified form but rather several interacting ones, typically set by or corresponding to different countries (compare, for example, U.S. vs British English). Other notable examples of pluricentric languages include French, Dutch, Korean, and Hebrew. The pluricentric nature of English as well as its worldwide pervasiveness are reflected in the fact there are not only vocabulary, grammar and spelling differences, but also many accents and dialects of English used in different countries and regions. In general, these all remain mutually intelligible, with a few exceptions around the extremes and in relation to idioms and slang.

[3] https://www.ethnologue.com/guides/ethnologue200

[4] https://www.ethnologue.com/guides/most-spoken-languages

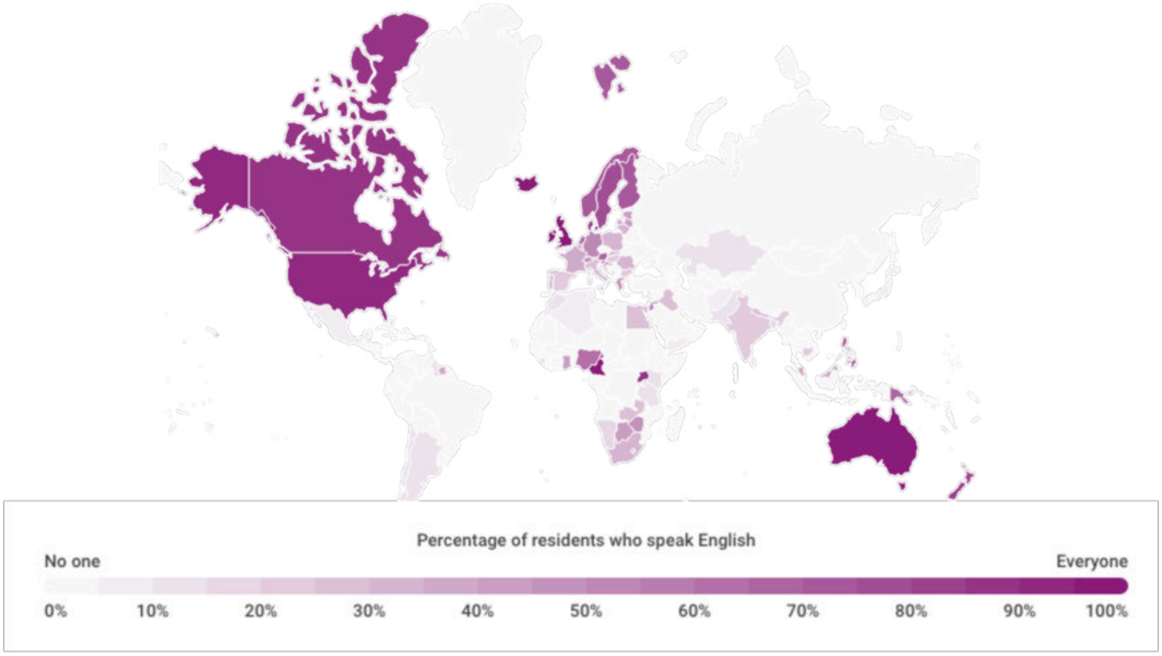

Figure 1: Percentage of residents per country around the world who speak English

English is an Indo-European language, belonging to the West Germanic group of Germanic languages (Bammesberger, 1992), and was historically spoken by the Anglo-Saxons. It is most closely related to Frisian and Low Saxon, while its vocabulary has been significantly influenced by other Germanic languages, particularly Old Norse, as well as Old Norman, French and Latin. It has evolved considerably over the last 1400 years: Old English originated from Germanic tribes along the Frisian North Sea coast, whose languages gradually separated into the Anglic languages in the British Isles and into the Frisian and Low German/Low Saxon languages elsewhere. Old English then evolved into Middle English and later Modern English (Robinson, 1992). Certain dialects of Old and Middle English evolved into other Anglic languages including Scots (Romaine, 1982) and various Irish dialects.[5] (Barry, 1982)

English shares a number of features of other Germanic languages, in particular Dutch, German, and Swedish (Durrell, 2006), sharing common roots from Proto-Germanic. It also shares features with Frisian, classifying it as an AngloFrisian language.

English uses the Latin alphabet with a left-to-right writing system. As an Indo-European language, it uses the ISO-639-1 code, with the two letter code *en*.

2.2 English in the Digital Sphere

Unsurprisingly, given its status as an international language, English is the most commonly used language online, representing approximately 60.4% of the top 10 million websites.[6] In comparison, China has the most internet users in the world but only 1.4% of the top 10M websites use Chinese. The Russian language is the second most widely used on the internet, representing approximately 8.5% of the top 10 million websites.[7]

[5] Though Irish (Gaelic) itself is derived from Celtic languages

[6] https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/

[7] Data as of January 2021 as reported by Visual Capitalist. “Share of speaking population” refers to the percentage of the global population that identifies as a speaker of a particular language, either as their native tongue or a

As of 31 March 2020, the internet was estimated to have approximately 1.186 billion English speaking users, accounting for 25.9% of all internet users around the world.[8] In comparison, the second highest figure is for Chinese, which had 888,453,068 internet users and a 19.4% share of all internet users. Meanwhile, although Russian is the second most widely used language on the internet, it represents only around 100 million users. It has been estimated that out of the 4.5 billion internet users, 18% of them are native English speakers, as of mid-2020.[9] The only language with a higher percentage of native speakers on the Internet is Chinese with 19% (i. e. unsurprisingly, almost all Chinese on the internet comes from native speakers). In terms of internet penetration for the English language, out of the 1.531 billion English speakers estimated for 2021 according to the Internet World Stats, 77.5% of them are internet users. In comparison, German and Japanese share the highest internet penetration rate out of the top 10 most widely used languages on the internet, each at 93.8%. Looking at internet user growth per language, we see that the number of English-speaking users has enjoyed a relatively modest growth rate of 742.9% in the last 20 years, compared with Arabic at 9,348%. In fact, out of the top 10 most widely used languages on the internet, English has the 3rd lowest growth rate in terms of users, only superior to that of German-speaking (236.2%) and Japanese-speaking users (152%).

3 What is Language Technology?

Natural language[10] 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 process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialized field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, Language Technology (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 fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably AI), mathematics and psychology among others. In practice, these communities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

Language Technology is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky et al., 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and

second language. “Percentage of 10 million websites” is based on traffic rankings from Alexa.com

[8] According to Internet World Stats https://www.internetworldstats.com/stats7.htm

[9] https://globalbydesign-com.cdn.ampproject.org/c/s/globalbydesign.com/2021/01/12/how-many-languages-should-your-website-support-2/amp/

[10] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

advances in machine learning, rule-based systems have been displaced by data-based ones, i. e. systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

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

LT is trying to provide solutions for the following main application areas:

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

LT is already fused in our everyday lives. As individual users we may be using it without even realizing it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name a few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[11]

4 Language Technology for English

While there has been an increasing interest in developing both data and tools for multilingual language processing in the last 20 years, as witnessed by the topics of long-standing shared tasks such as CONLL,[12] nevertheless English continues to be overwhelmingly dominant in every aspect of language processing. This is partially as a result of the dominance of the use and status of English in the digital sphere and as an international language, but also a circular problem related to the availability of existing low-level language processing tools and training data which provide an easy starting point for further development. It would be impossible to list all available tools and resources here, but we summarise the key findings of our survey based on the combined collection of ELE and ELG resources, which acts as a baseline for comparison of tools and resources for other languages. Due to the sheer number of resources available for English, it is hard to obtain any accurate figures for quantity since there is no single place where they are collected, and many are simply hosted on individual websites.

4.1 Language Data

Corpora

Unsurprisingly, thousands of corpora are freely available for English. The majority of these are covered by a Creative Commons license, although they may come with restrictions (e. g. attribution or no commercial use). Some are covered by shared task participation agreements, implying that they are freely available at least to task participants. A number of the corpora are released under licenses controlled by ELRA and thus only available to ELRA members. The LDC contains around 350 corpora for English for text alone, with several hundred more that do not specifically mention being text-based. The LDC grows by around 30 to 35 new corpora each year, and while these do not all include English, it does mean that new resources with contemporary language use appear with reasonable regularity. Corpora rarely go “out-of-date”, so even old corpora are still useful.

Within ELG, 255 monolingual English text corpora are available, most of which are described as being annotated, though the type and extent of annotation may vary widely (and is not recorded). Hugging Face also lists 309 monolingual English datasets, most of which are text based. ELG also covers 1164 multilingual corpora where one of the languages is English, although almost all of these are only bilingual. Hugging Face lists 106 multilingual datasets

[11] 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 an annual rate of 18,4% from 2020 to 2028 (https://tinyurl.com/2p9ed6tp). A different 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 an annual rate of 10,3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

[12] https://www.conll.org/previous-tasks

which include an English portion. Only two multimodal corpora are listed in ELG, containing text and video, and in one case also audio, but there are many more available elsewhere. The LDC catalogue lists a further 35 English corpora which contain both text and sound, 4 containing text and video, and 1 containing both text and images.

Lexical/conceptual Resources

ELG lists 35 monolingual lexical/conceptual resources, most of which are domain-specific, and includes also 3 ontologies. A further 13 English lexicons are listed in LDC, but it is likely that a huge number of freely available additional resources are available elsewhere, while the ELRA Universal catalogue lists 9 monolingual English lexicons. Additionally, 245 bilingual resources are listed, of which most are terminological resources and the rest lexicons. A further 96 multilingual resources are available from the LDC, of which again most are domain-specific. These all involve English as one of the languages. Many of all these resources are licensed via ELRA so only available to members. The ELRA Universal catalogue additionally lists 43 multilingual lexicons where English is one of the languages. Finally, 9 multimodal resources are listed resources (where text is one form), all of which are text/audio lexicons and are mostly concerned with pronunciation.

4.2 Language Technologies and Tools

English is very well-served generally by **spelling and grammar-checking tools**. The ELG repository contains three tools, but most operating systems have built-in spell-checking tools which can be used from the command line: for example, aspell and hunspell are common tools on linux. Most programming languages have at least one spell checking library: for example LanguageTool can be used from both Java and Python and also supports grammar checking. AllenNLP also provides models for reading comprehension.

Four **summarization system**s are available in ELG, although many more are available as open source or commercially in addition, including HuggingFace Transformers. Finally, 8 text-to-speech (TTS) systems are listed in ELG, although 7 of these are versions of MaryTTS with different voice models (3 female, and 4 male), while the 8th also supports Welsh. Again, there are a number of other open source TTS systems which could be integrated within ELG including: eSPeak-NG, gnuspeech, and the Festival Speech Synthesis System, as well as 62 TTS models hosted on HuggingFace which claim to cover English.

In terms of **models**, ELE lists 4,871 English models of which most are text related. Of these, 77 are for text summarization, 402 for translation, and 342 for various kinds of classification (including sentiment analysis), as well as 99 for token classification (including NER). The ELE collection does not contain any multimodal models for English, but Hugging Face provides some multimodal models as there are 62 Text-to-Speech models in the search results listed under English.

There are several major **infrastructures or toolkits** for language processing available, including GATE, Stanford CoreNLP, Stanford Stanza, NLTK, spaCy, Hugging Face Transformers, and OpenNLP, which all contain a variety of processing tools which can be used individually or as a collection. All of these frameworks support at least tokenisation, sentence splitting, PoS tagging, and Named Entity extraction, and some support many more tools such as sentiment analysis, or have specific support for domains such as medicine. While these toolkits were all primarily designed for English, they all have a number of models for other languages, and both this range and the functionality of these tools is continually improving.

As well as the main NLP toolkits there are a number of online services for NLP which provide many of the tools covered by this summary. For example, the big three (Google, Amazon, and Microsoft) all have offerings that cover the basic tools (tokenization, part of speech etc) as well as complex applications for named entity extraction, sentiment analysis etc. For more information on commercial and online offerings see Dale (2018) and Prinz et al. (2020); while ongoing research can be tracked on the NLP Progress website.[13]

Tools hosted as part of the ELG are all currently free to access, although calls may be subjected to a rate limit and or quota system. Access is through a common ELG REST API, making it trivial to call any of the services.

For low-level processing tasks, there are a few standalone tools and services contained within the ELG framework (5 tokenisers, one sentence splitter, 7 POS taggers), but many more are provided as part of standard APIs: for instance, Java provides a stringTokenizer class. Many other tools require these tasks as part of a large system and some allow access to these internal tools via the API, e. g. Keras (a python wrapper for TensorFlow) and Gensim (used for topic modelling) which provide access to their own internal tokenisers. Additionally, UDPipe is unique in that it provides an API for use with the statistical package R, rather than being command line-driven or providing bindings for Java or Python. It should be noted also that in general, tools for tokenisation and sentence splitting for European languages are more or less language-independent. POS tagging is also a reasonably well-solved problem for English (unlike a number of other languages) with most systems achieving a per token accuracy of 95% or above, although this is lower on text such as social media, for which there are now increasingly a number of models or specific taggers (e. g. GATE's TwitIE tagger). ACL curates a list of 19 PoS taggers, together with their performance on the Penn Treebank.[14]

In terms of **Information Extraction**, the ELG collection houses 20 Named Entity Recognition (NER) systems for English, of which roughly half are generic, with the rest being domain-specific, with domains/genres including biomedical, Twitter, dendrochronology, environment, chemistry and politics. However, all of the frameworks described above support some level of NER, ranging from simple classic NEs to focused entity types for specific applications. This is also an area which has seen many ML models released to tackle the task. For example, AllenNLP and Flair provide a number of models (hosted on Hugging Face).

English is also well served with systems which go beyond entity recognition to perform entity linking against general knowledge resources such as Wikidata, DBpedia, or YAGO as well as tools to link against domain specific resources such as SNOMED for clinical terms. A survey of approaches for linking against Wikidata is also available (Möller et al., 2021).

Tools which fall broadly into the **information retrieval** category cover a wide range of tasks, including question answering, which involves retrieving information from the web in order to answer a specific question. The ELG collection covers 25 cross-lingual IR systems of which 11 name English as one of the languages. The other language in this case is most frequently German, though other languages such as French and Italian are included. However, many of the remaining systems also enable search in a specified language but the ability to return results in other languages, including English. A further 2 monolingual information retrieval tools are listed for English. There are of course a number of commercial information retrieval engines available, both for generic and specialised tasks.

Similarly, there are 7 independent **sentiment analysis** systems available in ELG, although some of these are more language classification tools performing related tasks such as toxic and offensive language classifiers. There are a number of additional tools and models, mostly as part of larger NLP frameworks, since this is a field where ML approaches are popular, especially via HuggingFace.

Concerning **Machine Translation**, ELG repository lists 242 MT systems, of which 74 language pairs contains English as input and 67 as output, as summarised in Table 1. All EU official languages are covered in both directions, along with other languages from the EU (e. g. Catalan, Luxembourgish), accession candidates & EEA members (e. g. Norwegian Bok-

[13] https://nlpprogress.com

[14] https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)

mål & Nynorsk, Icelandic, Serbian) and other non-European languages (e. g., Arabic, Russian, Hindi). The most common pairing (regardless of direction) is English/German. Further services are expected to be integrated in the near future, notably from the NTEU project which provides translation models for *all* pairs of EU official languages; this will not broaden the language coverage in the English case but it will add an additional service in each direction between English and each of the other 23 EU official languages. There are additionally several freely available MT systems including: Moses,[15] Apertium,[16] OpenLogos, [17] and the Hugging Face Transformer MarianMT.[18]

	German	Finnish	Dutch	Swedish	Spanish	Czech	Danish	French	Polish	Romanian	Bulgarian	Croatian	Estonian	Hungarian	Latvian	Portuguese	Greek	Irish	Italian	Lithuanian	Maltese	Slovak	Slovenian	Others	Total
From English to…	6	3	2	4	2	3	2	2	3	2	2	1	2	1	2	2	1	1	1	1	1	1	1	28	74
To English from…	5	3	1	3	2	4	1	2	3	2	2	1	2	1	2	2	1	1	1	1	1	1	1	24	67

Table 1: MT resources and their language pairings available in ELG

4.3 Projects, Initiatives, Stakeholders

As part of the ELG project, and based on the META-NET Network of Excellence, 32 ELG National Competence Centres (NCCs) were established. Following META-NET's work to assess the situation of the LT landscape in Europe Rehm et al. (2016), the NCCs were asked to provide information about corresponding funding programmes, activities and challenges in their countries (Rehm et al., 2020). The EPSRC funds NLP actions, but this is primarily blue skies research. It is currently designated as a growth area, meaning that EPSRC has actively tried to increase the funding allocated to NLP research. However, LT and research infrastructures, in general, are not perceived as high funding priority (e. g., the UK is only an observing member of CLARIN).

At the same time, the University of Sheffield's GATE infrastructure (originally funded by EPSRC in the mid-1990s) and, most recently, GATE Cloud,[19] are among the most widely used and established LT tools, services, and platforms. There are several other major infrastructures and toolkits for English, described in the previous section. In September 2021, the UK's National Artificial Intelligence (AI) Strategy was launched with the aim of helping the UK 'strengthen its position as a global science superpower and seize the potential of modern technology to improve people's lives and solve global challenges such as climate change and public health' (Government, 2021). While the document emphasises the importance of AI for economic development, and notes the need to develop AI standards, ethics and infrastructure for the UK, there is scant mention of language-orientated AI, however.

In Ireland, NLP research is funded by SFI rather than EPSRC. Ireland also has a National AI Strategy entitled "AI – Here for Good", which aims to provide a high-level direction to the design, development and adoption of AI in Ireland, and makes some reference to Language Technology, mainly focusing on English (as opposed to Irish).[20]

In terms of LT providers, we have identified 53 major industrial organisations in the UK, including players such as BBC News Labs, the JISC, and Oxford University Press, and 246

[15] https://www.statmt.org/moses/
[16] https://wiki.apertium.org
[17] https://logos-os.dfki.de
[18] https://huggingface.co/docs/transformers/model_doc/marian
[19] https://cloud.gate.ac.uk
[20] https://www.gov.ie/en/publication/91f74-national-ai-strategy/ See page 42.

research groups or organisations based at 94 different universities. These research groups are split between various faculties and departments, comprising mostly Computer Science and Language departments, but also others such as Medicine, Architecture, Life Sciences and Education, Creative Industries, and Maths. In Ireland there are also extensive LT industry bodies and research centres (e. g. Apple, Accenture, Google, SoapBox Labs, AYLIEN, CeADAR, ADAPT Centre), whose primary focus is again mainly on supporting the English speaking rather than Irish speaking population.

5 Cross-Language Comparison

The LT field[21] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded unprecedented results. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[22] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing (e. g., facial expression recognition)
  - Human-computer interaction (e. g., tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

[21] This section has been provided by the editors.

[22] Tools tagged as “language independent” without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

5.2 Levels of Technology Support

We measured the relative technology support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[23]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[24] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[25]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific

[23] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e., 3%, 10% and 30%) were then used to define the bands per application area and resource type.

[24] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

[25] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

That said, we consider the current status of the ELG repository and the higher level findings below adequately representative with regard to the current existence of LT resources for Europe's languages.

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 2 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[26] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 2 visualises our findings.

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

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

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has

[26] In addition to the languages listed in Table 2, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

			Tools and Services							Language Resources					
			Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	**Overall**
EU official languages		Bulgarian	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Croatian	▒	░	░	▒	▒	▒	░	▒	░	▒	░	▒	▒
		Czech	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	░	▒	▒
		Danish	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	▒	▒	▒
		Dutch	▒	▓	▓	▒	▓	▓	▓	▒	░	▒	▒	▒	▒
		English	█	█	█	█	█	█	█	█	▓	█	▓	█	█
		Estonian	▒	▒	▓	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Finnish	▒	▒	▓	▒	▒	▓	▓	▒	▒	▒	▒	▒	▒
		French	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▒	▓	▓
		German	▓	█	█	█	█	█	█	▓	▓	▓	▒	▓	▓
		Greek	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	▒	▒	▒
		Hungarian	▒	▒	▒	▒	▒	▒	▓	▒	░	▒	▒	▒	▒
		Irish	░	░	░	░	░	▒	░	▒	░	▒	░	░	░
		Italian	▒	▓	▓	▓	▒	▓	▓	▒	▒	▒	░	▒	▒
		Latvian	▒	▒	▓	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Lithuanian	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	░	▒	▒
		Maltese	░	░	░	░	░	▒	░	░	░	▒	░	░	░
		Polish	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	░	▒	▒
		Portuguese	▓	▓	▓	▒	▒	▓	▓	▒	▒	▒	░	▒	▒
		Romanian	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	▒	▒	▒
		Slovak	▒	▒	▒	▒	▒	▒	▒	▒	░	▒	░	░	▒
		Slovenian	▒	▒	▒	▒	▒	▒	░	▒	░	▒	░	▒	▒
		Spanish	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓
		Swedish	▒	▒	▒	▒	▒	▓	▓	▒	▒	▒	░	▒	▒
(Co-)official languages	National level	Albanian	░	░	░	░	░	░	░	░	░	░	░	░	░
		Bosnian	░	░	░	░	░	░	░	░	░	░	░	░	░
		Icelandic	▒	▒	▒	░	░	▒	▒	░	░	▒	░	░	▒
		Luxembourgish	░	░	░	░	░	▒	░	░	░	░	░	░	░
		Macedonian	░	░	░	░	░	░	░	░	░	░	░	░	░
		Norwegian	▒	▒	▒	▒	▒	▓	▒	▒	░	▒	▒	░	▒
		Serbian	░	░	▒	▒	▒	▒	░	░	░	░	░	░	░
	Regional level	Basque	▒	▒	░	▒	░	▒	▒	░	░	░	░	░	▒
		Catalan	▒	▒	▒	▒	▒	▓	▒	▒	░	░	▒	▒	▒
		Faroese	░	░	░	░	░	░	░	░	░	░	░	░	░
		Frisian (Western)	░	░	░	░	░	░	░	░	░	░	░	░	░
		Galician	▒	░	▒	▒	░	▒	▒	░	░	░	░	░	▒
		Jerriais	░	░	░	░	░	░	░	░	░	░	░	░	░
		Low German	░	░	░	░	░	░	░	░	░	░	░	░	░
		Manx	░	░	░	░	░	░	░	░	░	░	░	░	░
		Mirandese	░	░	░	░	░	░	░	░	░	░	░	░	░
		Occitan	░	░	░	░	░	░	░	░	░	░	░	░	░
		Sorbian (Upper)	░	░	░	░	░	░	░	░	░	░	░	░	░
		Welsh	░	▒	░	░	░	▒	▒	░	░	░	▒	░	▒
All other languages			░	░	░	░	░	░	░	░	░	░	░	░	░

Table 2: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

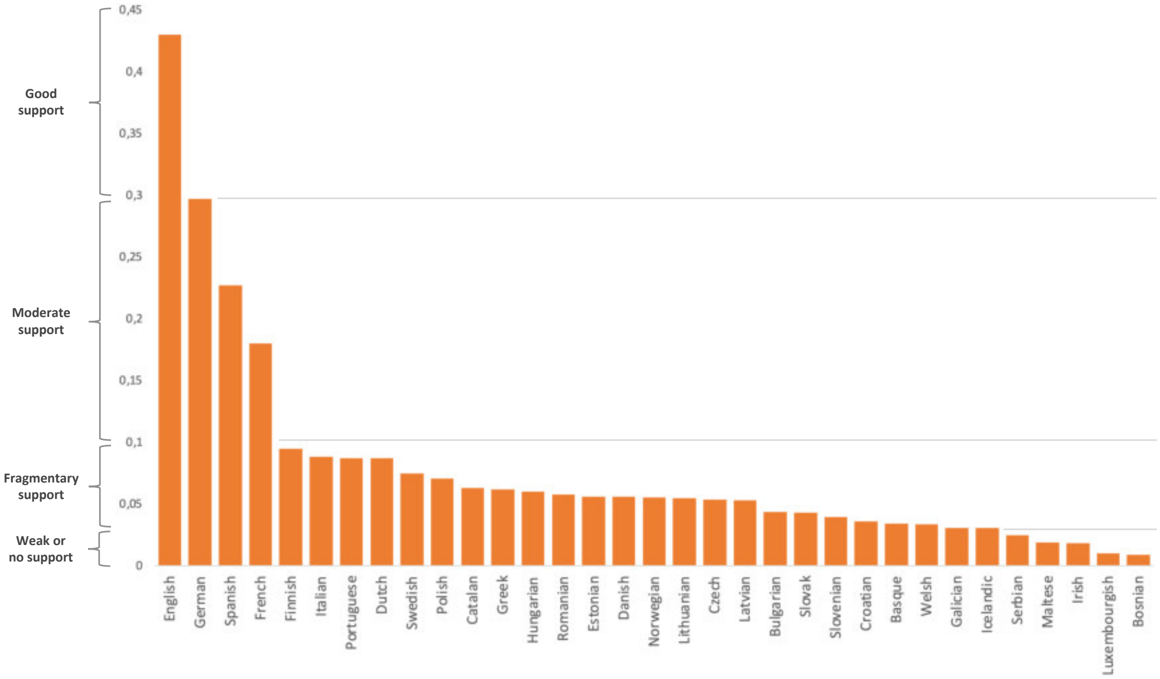

Figure 2: Overall state of technology support for selected European languages (2022)

significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally supported ones is still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality and avert the risks of digital extinction.

6 Summary and Conclusions

In summary, English is extremely well supported by Language Technology, which is unsurprising given its status in the digital world. Almost every tool and infrastructure or toolkit is first developed to handle English before being applied to other languages. Similarly, an enormous amount of data is available for English. These two factors have a circular effect: due to the amount of data available, training and testing new tools is much easier for English than other languages, and this leads to new models, tools, and resources being developed. The frequency with which English is used for online communication also provides a wealth of data from which to create new corpora, and the availability of a wide range of tools also makes it easier to annotate these with linguistic information. As tools improve, the accuracy and usefulness of pre-annotated corpora also improves, thereby making further tool development easier.

On the one hand, this is an excellent situation for those working on English data, and given the widespread use of English in the digital world, the usefulness of new tools is clear. On the other hand, this can be a double-edged sword for the development of language technology and resources for other languages. The availability of data, tools and resources for English has fed the enormous success of neural models for developing language technology applications, but the lack of data for other languages means that such deep learning models trained on English are not directly applicable. Recently, however, advances have been made in the development of cross-lingual transfer learning (CLTL) in order to build NLP models for a low-resource target language by leveraging labelled data from languages such as English with a high level of resources, or via a staged process whereby training data from English feeds the development of languages with moderate resources, which may have greater similarity to low-resource languages and can feed a further transfer process. Additionally, multilingual transfer settings enable training data in multiple source languages to be leveraged to further boost performance of low-resource languages. On the negative side, almost all languages are inevitably playing "catch-up" compared with English, and as can be seen from our survey, the differences in tools and resources available for European languages are striking. It is hard even to grasp a sense of how much is available for English, since resources are so disparate, and the figures reported in the collections of ELG, ELRA and other repositories are only the tip of the iceberg, as it is literally impossible to count every resource for English.

Improving the language situation for English is thus not a major issue to be resolved as such, but rather, how the development of English tools and resources can best be geared towards improving the situation of other European languages. However, the development of domain-specific resources for English is nevertheless quite patchy. Domains such as medicine and social media are relatively well-resourced, while recent advances in technology for legal AI, for instance, are improving areas traditionally slow to adopt technology. Surprisingly, perhaps, tools for use in many humanities domains are still lacking sophistication, despite long-standing research in digital archives, for instance. These shortcomings can only really be addressed by greater interaction with the language technology community.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

Alfred Bammesberger. The place of english in germanic and indo-european. *The Cambridge History of the English Language I: the beginnings to*, 1066:26–66, 1992.

Michael V. Barry. English in Ireland. In Richard W. Bailey and Manfred Görlach, editors, *English as a World Language*, pages 84–134. University of Michigan Press, 1982.

Noam Chomsky et al. Syntactic structures (the hague: Mouton, 1957). *Review of Verbal Behavior by BF Skinner, Language*, 35:26–58, 1957.

Michael Clyne. Pluricentric languages. *Different Norms in Different Nations, Mouton de Gruyter*, 1992.

Robert Dale. Text analytics apis, part 1: The bigger players. *Natural Language Engineering*, 24(2):317–324, 2018. doi: 10.1017/S1351324918000013.

M. Durrell. Germanic languages. 2006.

Cedric Möller, Jens Lehmann, and Ricardo Usbeck. Survey on English entity linking on wikidata. *arXiv preprint arXiv:2112.01989*, 2021.

Katja Prinz, Gerhard Backfried, Alexander Oberkersch, and Erinc Dikici. European language grid: D7.3 marketplace report, 2020. URL https://www.european-language-grid.eu/wp-content/uploads/2021/02/ELG-Deliverable-D7.3-final.pdf.

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

Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon. The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3315–3325, Marseille, France, 5 2020. European Language Resources Association (ELRA).

Orrin W Robinson. *Old English and its closest relatives: a survey of the earliest Germanic languages*. Stanford University Press, 1992.

S. Romaine. English language in Scotland. In Richard W. Bailey and Manfred Görlach, editors, *English as a World Language*, pages 51–70. University of Michigan Press, 1982.

Jan Svartvik and Geoffrey Leech. *One tongue: Many voices*. Springer, 2006.

Alan Turing. Computing machinery and intelligence in "mind", 1950.