EUROPEAN
LANGUAGE
EQUALITY

D1.21

Report on the Italian Language

Authors	Bernardo Magnini, Alberto Lavelli, Manuela Speranza
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	D1.21
Deliverable title	Report on the Italian Language
Type	Report
Number of pages	24
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	Bernardo Magnini, Alberto Lavelli, Manuela Speranza
Reviewers	Federico Gaspari, Jaroslava Hlavacova
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
AILC	Associazione Italiana per la Linguistica Computazionale
CL	Computational Linguistics
CLiC-it	Italian Conference on Computational Linguistcs
CLARIN	Common Language Resources and Technology Infrastructure
CNR	Consiglio Nazionale delle Ricerche (National Research Council)
COMPOSES	Compositional Operations in Semantic Space (ERC project, 2011-2016)
DL	Deep Learning
DLE	Digital Language Equality
DH	Digital Humanities
E3C	European Clinical Case Corpus (ELG pilot project, 2020-2021)
EC	European Commission
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
EU	European Union
EVALITA	Evaluation of NLP and Speech Tools for Italian
EVALITA4ELG	Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools for the ELG platform (ELG pilot project, 2020-2021)
GPU	Graphics Processing Unit
ISDT	Italian Stanford Dependency Treebank
ISST	Italian Syntactic-Semantic Treebank
ISTAT	Istituto nazionale di Statistica
LR	Language Resource/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MOUSSE	Multilingual, Open-text Unified Syntax-independent SEmantics (ERC project, 2017-2023)
MT	Machine Translation
MULTIJEDI	Multilingual Joint Word Sense Disambiguation (ERC project, 2011-2016)

NLG	Natural Language Generation
NLP	Natural Language Processing
SIL	Summer Institute of Linguistics
SR	Speaker Recognition
SSH	Social Sciences and the Humanities
TUT	Turin University Treebank
VIT	Venice Italian Treebank
WSD	Word Sense Disambiguation

Abstract

This report is an update of the META-NET white paper on "The Italian Language in the Digital Era" published in 2012. An update is very timely because of the many changes brought by neural approaches on language technologies (LTs) for the Italian language: not only the state of the art in almost all Natural Language Processing (NLP) tasks has been surpassed in the last ten years, furthermore, maybe most important, new tasks and new language resources (e. g. large pre-trained language models) have become central in everyday LT practice. While such changes are to a large extent common to other European languages, the Italian LT community has taken further advantage by three important factors: (i) the Associazione Italiana di Linguistica Computazionale (Italian Association for Computational Linguistics, AILC) founded in 2015 with the goal of establishing common ground for the Italian LT community; (ii) CLiC-it, the annual Italian Conference on Computational Linguistics, now the most important forum for computational linguistics in Italy; (iii) the EVALITA (Evaluation of NLP and Speech Tools for Italian) evaluation campaigns, which provide a shared framework where different systems and approaches can be evaluated in a consistent matter. Given this rather favorable context (new neural approaches and strong community initiatives), we are seeing a widespread expansion of interest in LT for Italian, both in the academia and in industry. At the same time, the LT bar is continuously moved upwards, which requires adequate efforts and investments. This is particularly needed in areas, such as for instance dialogue systems, where Italian is still lacking sufficient language resources, and in application domains, such as biomedicine, where progress is still limited.

Riassunto

Questo documento costituisce un aggiornamento rispetto al libro bianco META-NET su "La lingua italiana nell'era digitale" pubblicato nel 2012 e arriva al momento opportuno se prendiamo in considerazione i numerosi cambiamenti provocati dagli approcci neurali nel campo delle tecnologie del linguaggio (LT) per la lingua italiana: non solo negli ultimi dieci anni lo Stato dell'arte è progredito in quasi tutti i compiti di Natural Language Processing (NLP), ma sono anche diventati centrali, nella pratica quotidiana delle LT, nuovi compiti e nuove risorse linguistiche (ad esempio, grandi modelli del linguaggio pre-addestrati), il che è forse ancora più significativo.

Tali cambiamenti sono in larga misura comuni ad altre lingue europee, ma occorre osservare che la comunità italiana ha tratto ulteriore vantaggio da tre importanti fattori verificatisi negli ultimi anni: (i) l'Associazione Italiana di Linguistica Computazionale (AILC), fondata nel 2015 con l'obiettivo di stabilire un terreno comune per la comunità italiana delle LT; (ii) CLiC-it, l'annuale Conferenza Italiana di Linguistica Computazionale, che ad oggi costituisce il più importante forum per la linguistica computazionale in Italia; (iii) le campagne di valutazione EVALITA (Valutazione di Strumenti di NLP e di elaborazione del parlato per l'italiano), che forniscono un quadro comune dove sistemi e approcci diversi possono essere valutati in modo coerente.

In questo contesto particolarmente favorevole (i nuovi approcci neurali e le forti iniziative della comunità), l'interesse per le LT per l'italiano si sta diffondendo in maniera evidente, sia nel mondo accademico che nell'industria. Allo stesso tempo, l'asticella delle LT viene continuamente spostata verso l'alto, il che richiede sforzi e investimenti adeguati. Questo è necessario specialmente in settori dove l'italiano è ancora carente di risorse linguistiche (per esempio, l'area dei sistemi di dialogo) e in domini applicativi in cui i progressi sono ancora limitati (per esempio la biomedicina).

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 it 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 and 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, 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. 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 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 Italian Language in the Digital Age

2.1 General Facts

Italian is the de facto official language of Italy (it does not formally appear in the Italian Constitution as the official language, although it has been considered the official language at least since the reunification of Italy) and San Marino, where it is by far the most widely spoken language and almost all media (television, newspapers, movies, etc.) are produced in Italian. However, other languages are co-official within certain regions, including French in Val d'Aosta, German in Trentino-Alto Adige, and Sardinian in Sardinia.

In Switzerland, Italian is one of four official languages, spoken mainly in Canton Grigioni and Canton Ticino. In the Vatican City State, it is one of the official languages (all laws and regulations of the state are published in Italian). Moreover, it is the main working language of the Holy See, serving as the common language in the Roman Catholic hierarchy as well as the official language of the Sovereign Military Order of Malta. It has an official minority status in Slovenia and in Croatia.

Italian formerly had official status in Albania, Malta, Monaco, Montenegro (Kotor), Greece (Ionian Islands and Dodecanese). It used to be an official language in the former colonies of Italy in East Africa and North Africa, where it still has a significant role in various sectors. Italian is included under the languages covered by the European Charter for Regional or Minority languages in Bosnia and Herzegovina and in Romania, although it is neither a co-official nor a protected language in these countries. Italian is also spoken by very large immigrant and expatriate communities in the Americas and Australia.

Italian is a major European language, being one of the official languages of the Organization for Security and Co-operation in Europe and one of the working languages of the Council of Europe. According to the outcome of a 2012 survey (Lan, 2012), Italian is the native language of around 15% of the EU population and thus the second most widely spoken language after German (Keating, 2020). According to the 2019 edition of Ethnologue (Eth, 2021), a language reference published by SIL International, Italian has 64.8 million first language

[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.

speakers which makes it the 22nd most spoken language in the world (0.842% of the world population).

Around 56 million native speakers of Italian reside in Italy; it has been estimated (Calzolari et al., 2012) that, additionally, 470,000 first language speakers of Italian reside in Switzerland, 280,000 in Belgium, 70,000 in Croatia, 1,000,000 in France, 548,000 in Germany, 21,000 in Luxembourg, 27,000 in Malta, 2,500 in Romania, 4,000 in Slovenia and 200,000 in the United Kingdom. Very large immigrant communities, each consisting of over 500,000 people still speaking Italian, are found in Argentina, Brazil, Canada and the United States (Calzolari et al., 2012).

Italian belongs to the Indo-European language family of the Romance languages descending from the vulgar Latin; among the Romance national languages, Italian is the closest to Latin in terms of vocabulary. It has a 7 vowel sound system and almost all native Italian words end with vowels. Unlike most other Romance languages, it retains Latin's contrast between short and long consonants. The writing system is close to being a phonemic orthography, i. e. Italian has very regular spelling with a largely systematic and predictable correspondence between letters and sounds.

Italian grammar is typical of Romance languages in general. Cases exist for pronouns (nominative, accusative, dative), but not for nouns. There are two genders (masculine and feminine). Nouns, adjectives, and articles inflect for gender and number (singular and plural). Subject pronouns are usually dropped, their presence implied by verbal inflections. There are numerous contractions of prepositions with subsequent articles and numerous productive suffixes for diminutive, augmentative, pejorative, attenuating, etc.

A peculiar characteristic of Italian is that many native speakers of Italian residing in Italy are actually native bilingual speakers of both Italian (either in its standard form or regional varieties) and of an Italian dialect. Most specifically, according to a 2015 ISTAT report, 14% of the Italian population above the age of 6 (around 8 million people) uses mainly an Italian dialect at home and 32.2% use both Italian and an Italian dialect (IST, 2015). The Italian dialects may differ significantly from Italian (in fact, some are distinct enough to be considered as separate languages) and they played a significant role in the development of regional Italian variants, mainly with respect to the prosody, phonetics and lexicon.

A key institution of recognised authority for research on the Italian language, also in relation to its regional varieties, is the "Accademia della Crusca",[2] founded in Florence in the second half of the 16th century. Its main early accomplishment was the "Vocabolario degli Accademici della Crusca" (1612), the first dictionary of the Italian language. At present, its activity is centered on acquiring and spreading not only historical knowledge of the Italian language, but also awareness of the present evolution of Italian in the era of the information society.

A similar role is played by Società Dante Alighieri,[3] which has a key and highly regarded function in protecting, promoting and disseminating Italian culture, in particular abroad.

Since the 1950s, the presence of the American way of life in the media strongly influenced Italian culture and language. Due to the continuing triumph of English-language music since the 1960s, English acquired the status of a 'cool' language, which is reflected in the large number of present-day loan words from English (so-called anglicisms).

Partially as a reaction to the increasing presence of anglicisms in Italian, since 2001 a number of proposals have been submitted to the Italian Parliament to create the "Consiglio superiore della lingua italiana" (CSLI – High Council for the Italian Language), with the aim of protecting, promoting and disseminating Italian culture, particularly through initiatives encouraging the correct use of the Italian language, specifically in schools, communication media and commerce.

[2] https://accademiadellacrusca.it

[3] https://ladante.it

2.2 Italian in the Digital Sphere

The Digital Report is a survey that was conducted in 2020 by “We Are Social” in collaboration with Hootsuite, with the aim of collecting data on the use of internet and social platforms both at the global and local level (Dig, 2021). It reports that in Italy there were over 1 million people connected to internet for the first time in 2020, for a total of over 50 million internet users.

As far as the average time spent online by internet users aged 16 to 64, the Digital Report says that it amounts to over six hours per day, of which over three are spent watching television (broadcast and streaming) and almost two using social media; a considerable amount of time is also spent reading press media, listening to music streaming services and broadcast radio.

Social media is becoming more and more popular in Italy, with 98.5% of internet users who visited or used at least one social network or a messaging service in a month. The percentage of internet users who actively engaged with or contributed to social media in the same period amounts to 85.2% and every internet user has on average 7.8 social media accounts.

Italian is at fourteenth place in the ranking of the most used languages on the internet, as W3Techs estimates it to be used by 0.7% of the top 10 million websites on the World Wide Web (Use, 2021).

3 What is Language Technology?

Natural language[4] 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 specialised 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, 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 could be automatically analysed and/or produced. Gradually, with the evolution and 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

[4] 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).

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 revolutionising 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 realising 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.[5]

4 Language Technology Support for Italian

The resources and tools mentioned in this section have been taken from different sources. First of all, there are those already available in the European Language Grid (ELG) catalogue[6] in October 2021 (a limited number). Then, others have been identified mining previous repositories of linguistic resources and tools, such as CLARIN[7] (Hinrichs and Krauwer, 2014; Jong et al., 2018) and LREMap.[8] However, caution is required when attempting to automatically extract resources from some repositories. For example, on the one hand, CLARIN's goal is somehow different (and much wider) than ELE's and so a manual check is needed to assess the relevance of the resources. On the other hand, some LREMap entries contain very partial information and occasionally some of the resources do not appear to be available anywhere.

4.1 Language Data

A considerable part of the publicly available language resources for Italian have been produced in a number of EVALITA evaluation campaigns.[9] In this context, it is worth mentioning the EVALITA4ELG[10] project, the goal of which is to enable the ELG users to access the resources and models for the Italian language produced over the years in the context of the EVALITA evaluation campaign. The aim is to build the catalogue of EVALITA resources and tasks ranging from traditional tasks like POS-tagging and parsing to recent and popular ones such as sentiment analysis and hate speech detection on social media, and integrate them into the ELG platform. The project includes the integration of state-of-the-art LT services into the ELG platform, accessible as web services (Patti et al., 2020).

Monolingual text corpora

Examples of large-scale corpora for Italian are the following: CORIS/CODIS (150 million words) (Rossini Favretti et al., 2002), itWaC (1.6 billion words, texts downloaded from the Web and slightly cleaned up) (Baroni et al., 2009), Twita (over 150 million tweets) (Basile and Nissim, 2013), PAISÀ (250 million tokens) (Lyding et al., 2014).

[5] 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://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). 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).

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

[7] http://www.clarin.eu and the Italian node http://www.clarin-it.it

[8] https://lremap.elra.info

[9] https://www.evalita.it

[10] ELG (European Language Grid) Pilot Projects Open Call 1 (Grant Agreement No. 825627 – H2020, ICT 2018-2020 FSTP) http://evalita4elg.di.unito.it

In the context of the EVALITA evaluation campaign 62 tasks (with the availability of corresponding annotated data) have been organised. The tasks range from lemmatisation to sentiment analysis, covering both written texts and speech tools. In the first EVALITA editions, tasks have mainly used balanced or general reference corpora (all corpora of the written language, frequently including newspaper articles or being a balanced sample of different genres). Since 2009, data from computer-mediated communications started being used. Wikipedia articles were collected to prepare the dataset for the POS tagging and the textual entailment task in 2009, whereas in 2011 this genre was used in the parsing and super sense tagging challenges. From 2014, social media data started being used (in most cases, extracted from Twitter). In the majority of EVALITA tasks the focus has been on written textual data of various genres (only 13 out of 62 tasks make use of speech data). In 2014, speech tasks have been more numerous than textual ones (4 our of 7), whereas in the first and latest editions there were no tasks that focused on speech data.

Treebanks based on different formalisms are available for Italian (listed in reverse chronological order): the Italian Universal Dependencies treebank (278,429 tokens), the Italian Stanford Dependency Treebank (ISDT), the Turin University Treebank[11] (TUT, 101,309 tokens) (Bosco et al., 2000), the Italian Syntactic–Semantic Treebank (ISST, 80,967 tokens) and the Venice Italian Treebank[12] (VIT, 272,000 words).

Launched at first to develop a missing resource for Italian (i. e. the Turin University Treebank), the TUT project[13] eventually released other data sets: corpora for Sentiment Analysis (SentiTUT, Felicittà, and labuonascuola). Datasets for various tasks (such as COVID-19 emergency, hate speech detection, estimation of the degree of happiness in Italian cities) have been made available in the context of TWITA,[14] a collection of tweets identified as being written in the Italian language.

Bi- and multi-lingual text corpora

As for multilingual corpora, there are first of all the corpora related to MT. e. g. the Europarl Parallel Corpus[15] (up to around 60 million words per language) and the COVID-19 EUROPARL v2 dataset. Bilingual (EN-IT) Corpus[16] (650 Translation Units).

Another multilingual corpus is WIT^3 (Web Inventory of Transcribed and Translated Talks)[17] that contains a collection of transcribed and translated talks. The core of the dataset is from the Ted Talks corpus. Concerning Italian, the size of the parallel corpora for the various pairs of languages ranges between 410,000 and 1,600,000 words.

IWSLT 2017 Human Post-Editing data. The human evaluation (HE) dataset created for Dutch to German (NlDe) and Romanian to Italian (RoIt) MT tasks was a subset of the official test set of the IWSLT 2017 evaluation campaign.

ParTUT (Bosco et al., 2012) is a project for the development of a multilingual parallel treebank for Italian, English and French. The aim is twofold: building an aligned parallel treebank for Italian, English and French, by applying and extending the Italian treebank schema to other languages, and studying how the schema can be used to address issues typically related to parallel corpora.

The E3C project[18] (European Clinical Case Corpus; (Magnini et al., 2020)) aims at the creation of a corpus of clinical cases in 5 European languages (Italian, English, Spanish, French,

[11] http://www.di.unito.it/~tutreeb/

[12] http://catalog.elra.info/en-us/repository/browse/ELRA-W0040/

[13] http://www.di.unito.it/~tutreeb/

[14] http://twita.di.unito.it

[15] https://www.statmt.org/europarl/

[16] https://elrc-share.eu/repository/browse/covid-19-europarl-v2-dataset-bilingual-en-it/b046a370941f11ea913100155d026706cad48825217445eab849573cc66f9448/

[17] https://wit3.fbk.eu

[18] ELG (European Language Grid) Pilot Projects Open Call 1 (Grant Agreement No. 825627 – H2020, ICT 2018-2020

and Basque). The corpus is annotated with clinical entities (e. g. symptoms and pathologies), temporal information, and factuality.

Multimodal corpora (audio, video)

First of all, there are the resources made available in EVALITA (Patti et al., 2020).

Then, there are the multilingual audio corpora used as standard benchmarks in spoken language translation: e. g. EPIC (European Parliament Interpretation Corpus) (Russo et al., 2012) and Europarl-ST.[19]

Lexical/conceptual resources

Concerning lexical/conceptual resources, the following are worth mentioning because of their multilingual approach:

- BabelNet (Navigli et al., 2021), a multilingual encyclopedic dictionary, with wide lexicographic and encyclopedic coverage of terms, and a semantic network/ontology which connects concepts and named entities in a very large network of semantic relations, made up of about 20 million entries.
- EuroWordnet,[20] a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian)
- MultiWordnet,[21] a multilingual lexical database in which the Italian WordNet is strictly aligned with Princeton WordNet 1.6

Models and grammars

Starting from BERT (Devlin et al., 2019), transformer-based models have pushed the state of the art in many areas of NLP. To help the researchers not to get lost in the BERT models, in Bert Lang Street[22] (Nozza et al., 2020) they have indexed 31 BERT-based models, 19 Languages and 28 Tasks (with a total of 178 entries). The dimensions taken into account are the language, the NLP task, the model, the domain, the performance, …. For Italian, 13 entries are listed.

4.2 Language Technologies and Tools

Different pipelines composed of a series of NLP modules are available for Italian:

- deepnl[23] (Deep Learning for Natural Language Processing),
- LinguA[24] (Linguistic Annotation pipeline),
- NLTK[25] (Natural Language Toolkit),
- spaCy[26] (Industrial-Strength Natural Language Processing in python),

FSTP). https://e3c.fbk.eu

[19] https://mllp.upv.es/europarl-st/

[20] http://www.ilc.cnr.it/it/content/eurowordnet

[21] https://multiwordnet.fbk.eu

[22] https://bertlang.unibocconi.it

[23] https://github.com/attardi/deepnl

[24] http://www.italianlp.it/demo/linguistic-annotation-tool/

[25] https://www.nltk.org

[26] https://spacy.io

- TextPro[27] (Text Processing Tools),
- Tint[28] (The Italian NLP Tool),

The items in the (non-exhaustive) list are different for architectures, technologies (traditional machine learning, deep learning, ...), programming languages (java, python, ...), licenses, and witness the relative wealth of publicly available modules for NLP tasks on Italian.

4.3 Projects, Initiatives, Stakeholders

National funding programmes

The last LT funding programme in Italy dates back to 1999–2001. Since then, there has been no specific programme, nor is one foreseen in the near future. The National Programme for Research (2015–2020) identifies four technological clusters grouping 12 thematic areas – language is never mentioned. Italy lacks a coordinated plan for the development of LT. The Italian NLP sector finds some financial support through the national funding provided by the Ministry for University and Research (MIUR): in 2017, 111 million EUR out of a total budget of 391 million EUR were allocated to the wider SSH sector, where LT activities can receive some support (but LT is not recognised as a specific sector). At the regional level, programmes such as the Working Regional Programme from the European Fund for Regional Development (2014–2020) provide support to the wider ICT sector for the development of new technologies and innovation. Thus, funding is, in principle, available depending on the initiative and capacity of individual researchers and groups, but Italy severely lacks a coordinated research and development framework.

National research infrastructures.

There is no national research infrastructure dedicated to LTs in Italy. Corpora (both annotated and not annotated, benchmarks, tools for several tasks) for the Italian language are, however, available either through web sites of single research institutions, or through shared infrastructures at the European level, including the CLARIN repository and the European Language Grid repository.

Important projects

In order to provide concrete instances of relevant projects on LT, here we mention three European Research Council (ERC) projects granted to Italian institutions that are related to Language Technologies.

COMPOSES (Compositional Operations in Semantic Space) (Baroni et al., 2014) addresses the ability to construct new meanings by combining words into larger constituents, as one of the fundamental and peculiarly human characteristics of language.

MULTIJEDI[29] (Multilingual Joint Word Sense Disambiguation) proposes a research program that investigates radically new directions for performing multilingual word sense disambiguation (WSD). The key intuition underlying the project is that WSD can be performed globally to exploit at the same time knowledge available in many languages.

[27] https://textpro.fbk.eu
[28] https://dh.fbk.eu/research/tint/
[29] http://lcl.uniroma1.it/multijedi/

MOUSSE[30] (Multilingual, Open-text Unified Syntax-independent SEmantics) tackles the long-lasting challenge in NLP of semantic parsing, which has recently gained popularity. The MOUSSE project substantially contributed to the development of BabelNet Navigli et al. (2021).

Relevant scientific initiatives

Despite the lack of national funding programmes, the Italian community is rather active at the international level. Just to mention few events, Italy hosted EACL 2006 (the 11th Conference of the European Chapter of the Association for Computational Linguistics) in Trento, and ACL 2019 (the 57th Annual Meeting of the Association for Computational Linguistics) in Florence. Italian researchers have been chairing several LT conferences, including various editions of the Language Resources and Evaluation Conference (LREC), ACL 2021 (Program Chair) and ACL 2022 (General Chair).

In the last few years a series of initiatives have been taking place in the Italian NLP community. In 2007, the first edition of EVALITA[31] (Evaluation of NLP and Speech Tools for Italian) was held. The general objective of EVALITA is to promote the development of language and speech technologies for the Italian language, providing a shared framework where different systems and approaches can be evaluated in a consistent manner. The good response obtained by EVALITA, both in the number of participants and in the quality of results, showed that it is worth pursuing such goals for the Italian language. As a side effect of the evaluation campaign, both training and test data are available to the scientific community as benchmarks for future improvements (see Section 4.1 for more details). The first EVALITA edition was then followed by other six successful editions, the last of them in 2020.

Following the strong interest raised by EVALITA, the Associazione Italiana di Linguistica Computazionale[32] (Italian Association for Computational Linguistics, AILC) was founded in 2015, with the goal of establishing common ground for the Italian LT community, considering background and experiences both from the humanities and computer science.

A second relevant initiative on LT in Italy is CLiC-it, the annual Italian Conference on Computational Linguistics.[33] The first edition of CLiC-it was held in Pisa in 2014, which was followed by editions in Trento (2015), Naples (2016), Rome (2017), Turin (2018), Bari (2019), Bologna (2020) and Milan (2022). CLiC-it has become the most important forum for computational linguistics in Italy, and has obtained the important goal of stimulating the production of high-quality research and resources for the Italian language. Such efforts helped to shape the Italian landscape concerning NLP resources and tools.

Another relevant initiative concerning Italian (LTs) is the work carried on by the European Language Resource Coordination[34] (ELRC). One of the aims of the Italian ELRC is to mobilise public sector bodies to share their high-quality translated data on ELRC-SHARE, the repository used for documenting, storing, browsing and accessing Language Resources collected through the ELRC and considered useful for feeding the CEF Automated Translation (CEF.AT) platform. In June 2021 the third workshop of the Italian ELRC was the occasion for a discussion on the current situation and on the perspectives of language technologies for Italian.

Finally, it is worth to mention the Lectures on Computational Linguistics,[35] an AILC initiative targeting students (both master and PhD) and aiming at providing core competence in the LT field. The Lectures take place over three days, offering tutorials and labs by leading

[30] http://mousse-project.org
[31] https://www.evalita.it
[32] https://www.ai-lc.it
[33] https://www.ai-lc.it/en/conferences/clic-it/
[34] https://lr-coordination.eu
[35] https://www.ai-lc.it/lectures/

experts. With more than sixty participants per edition, and after five editions, the Lectures are now considered the major educational event for LT in Italy.

LT providers

The Italian LT academic community is relatively well distributed over the whole Italian territory. With a certain level of approximation, there are about 15 university departments in human sciences (e. g. linguistics, digital humanities, cognitive sciences) where research on computational linguistics is carried out, and about the same number of departments in computer science. In addition, there are departments of the National Research Council (CNR) and local research institutions, which are very active in the field of computational linguistics and NLP.

As for the industrial providers, in Italy there are more than one hundred companies that can be considered as active developers in the LT field, with a significant portion of companies that have appeared in the last few years, under the push of deep learning approaches and artificial intelligence. Among the LT companies, most are SMEs with connections with university departments, while only few of them also conduct business outside Italy.

5 Cross-Language Comparison

The LT field[36] 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 results never seen before. 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[37] 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)

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

[37] 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.

- 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.)

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 $\geq$3% and $<$10% of the ELG resources of the same type
3. **Moderate support**: the language is present in $\geq$10% and $<$30% of the ELG resources of the same type
4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[38]

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[39] 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

[38] 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.

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

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.[40]

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 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 1 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,[41] 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 1 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

[40] 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.

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

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

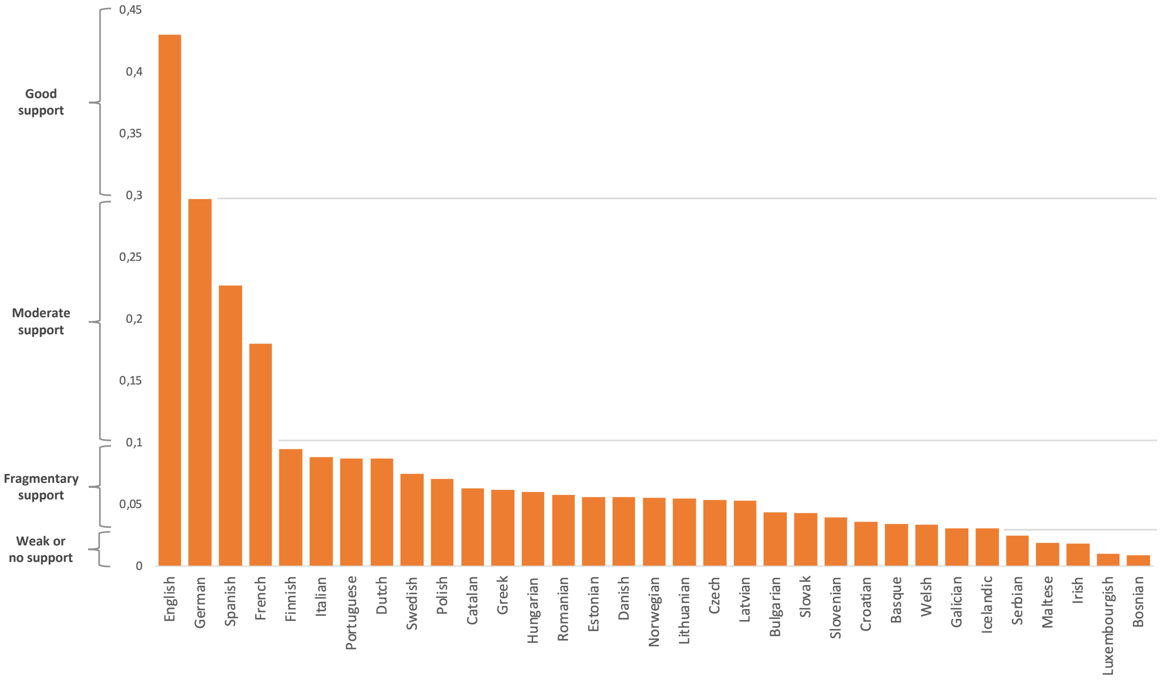

Figure 1: Overall state of technology support for selected European languages (2022)

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 the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally supported ones 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

Most of LT progresses in the last ten years (e. g. in machine translation, speech recognition, information extraction) are due to the introduction of powerful neural architectures, able to discover patterns in human language productions. While current research aims at diminishing the level of human supervision (for instance, using large self-trained language models), LTs are still largely based on supervised approaches, requiring massive amounts of annotated language resources. In addition, the LT bar is quickly moved upwards, with new tasks being frequently designed, and new resources being necessary for training. Within such a context, keeping up LT research and applications for Italian with the progresses of major languages, particularly English and Chinese, is very challenging.

In the report we have highlighted both strengths and weaknesses of LT for Italian. As for strengths, we mention the initiatives managed by the Italian Association for Computational Linguistics: the CLiC-it conference, the EVALITA evaluation campaigns, the Italian Journal of Computational Linguistics, and the annual seminars proposed by the Lectures of Computational Linguistics. Those initiatives are fundamental in order to motivate and grow an active community on LTs for Italian. Most of the LT resources and tools for Italian developed in the last ten years originate from such initiatives. Weaknesses are mainly due to the difficulty to keep up with the fast advances of English, which, in the best case means replicating English resources (e. g. by automatic translation), while in the worst case it means that many tasks and resources are simple not available for Italian. This is the case, for instance, of data-driven conversational agents, where annotated dialogues for Italian are limited to few datasets. Another limitation concerns the poor availability of domain specific resources for Italian, affecting the development of applications (e. g. in the medical domain).

In order to reduce, or not to further enlarge, the gap with English, few coordinated actions may help: (i) extend the LT community in Italy, particularly involving human science departments at universities; this action requires promotion and lobbying on LT at political level; (ii) promote the production, and re-use, of high quality LT resources to be used for data-driven approaches in NLP tasks, complementing voluntary based initiatives like EVALITA; this action requires substantial investments, mainly from the public sector. (iii) promote collaboration between academy and industry (not only in Italy) in few selected LT application domains, possibly in connection with emerging opportunities in the broader field of artificial intelligence; this action requires private-public investments (e. g. industry and innovation doctoral schools) and good connections with translational initiatives (e. g. ELE project).

References

Europeans and their languages. https://europa.eu/eurobarometer/api/archives/ebs/ebs_386_en.pdf, 2012. Archived 6 January 2016 at the Wayback Machine, https://web.archive.org/web/20160106183351/http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf.

L'uso della lingua italiana, dei dialetti e delle lingue straniere. https://www.istat.it/it/files/2017/12/Report_Uso-italiano_dialetti_altrelingue_2015.pdf, 2015. Accessed 12 November 2021.

Digital 2021 - I dati italiani. https://wearesocial.com/it/blog/2021/02/digital-2021-i-dati-italiani/, 2021. Accessed: 2021-10-30.

Ethnologue. summary by language size. https://www.ethnologue.com, 2021. Archived from the original on March 2019. Retrieved 12 November 2021.

Usage statistics of content languages for websites. https://w3techs.com/technologies/overview/content_language, 2021. archive.fo. Archived from the original on 12 November 2021. Accessed 12 November 2021.

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.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, September 2009. ISSN 1574-0218. doi: 10.1007/s10579-009-9081-4. URL https://doi.org/10.1007/s10579-009-9081-4.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. Frege in space: A program for composition distributional semantics. In *LILT*, 2014.

Valerio Basile and Malvina Nissim. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Leonardo Lesmo. Building a treebank for Italian: a data-driven annotation schema. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000. European Language Resources Association (ELRA).

Cristina Bosco, Manuela Sanguinetti, and Leonardo Lesmo. The Parallel-TUT: a multilingual and multiformat treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Nicoletta Calzolari, Bernardo Magnini, Claudia Soria, and Manuela Speranza. *La Lingua Italiana nell'Era Digitale – The Italian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

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

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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Erhard Hinrichs and Steven Krauwer. The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, May 2014. URL http://dspace.library.uu.nl/handle/1874/307981.

Franciska De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. Clarin: Towards fair and responsible data science using language resources. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélčne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Dave Keating. Despite Brexit, English remains the EU's most spoken language by far. *Forbes. Retrieved 7 February 2020.*, 2020.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-0406. URL https://aclanthology.org/W14-0406.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy, December 2020. Associazione Italiana di Linguistica Computazionale. URL http://ceur-ws.org/Vol-2769/paper_55.pdf.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. Ten years of BabelNet: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/620. URL https://doi.org/10.24963/ijcai.2021/620. Survey Track.

Debora Nozza, Federico Bianchi, and Dirk Hovy. What the [MASK]? making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*, 2020.

Viviana Patti, Valerio Basile, Cristina Bosco, Rossella Varvara, Michael Fell, Andrea Bolioli, and Alessio Bosca. EVALITA4ELG: Italian benchmark linguistic resources, NLP services and tools for the ELG platform. *Italian Journal of Computational Linguistics*, 6(2):105–129, 2020.

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.

Rema Rossini Favretti, Fabio Tamburini, and Cristiana De Santis. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. *A rainbow of corpora: Corpus linguistics and the languages of the world*, pages 27–38, 2002.

Mariachiara Russo, Claudio Bendazzoli, Annalisa Sandrelli, and Nicoletta Spinolo. The European parliament interpreting corpus (EPIC): Implementation and developments. 2012.

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.