EUROPEAN LANGUAGE EQUALITY

D1.25

Report on the Maltese Language

Authors	Mike Rosner, Claudia Borg
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	D1.25
Deliverable title	Report on the Maltese Language
Type	Report
Number of pages	35
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	Mike Rosner, Claudia Borg
Reviewers	Aritz Farwell, Annika Grützner-Zahn
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
ASR	Automatic Speech Recognition
CEF	Connecting Europe Facility
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CLEM	Corpus of Learner English in Malta
DAI	Department of Artificial Intelligence
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DL	Deep Learning
DLE	Digital Language Equality
EC	European Commission
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)
ELRC	European Language Resource Coordination
EMA	European Medicines Agency
EP	European Parliament
EU	European Union
FITA	Foundation for IT Awareness
GPU	Graphics Processing Unit
HPC	High-Performance Computing
ILLT	Institute of Linguistics and Language Technology (University of Malta)
IT	Information Technology
KMM	Korpus Malti Mitkellem
LR	Language Resources/Resources

LSM	Maltese Sign Language
LT	Language Technology/Technologies
MDIA	Malta Digital Innovation Authority
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
MITA	Malta Information Technology Agency
ML	Machine Learning
MLRS	Maltese Language Resource Server
MSA	Maltese Standards Authority
MSE	Maltese Speech Engine
MT	Machine Translation
MWE	Multi-Word Expressions
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
NLTP	National Language Technology Platform
PC	Personal Computer
POS	Part of Speech
R&D	Research and Development
SWOT	Strengths, Weaknesses, Opportunities, and Threats
TTS	Text-to-Speech
UM	University of Malta
YPP	YouTube Partnership Programme

Abstract

This report is an update to *The Maltese Language the Digital Age* (Rosner and Joachimsen, 2012b), a previous report that formed part of the META-NET White Paper Series on Europe's languages. The present report, which also forms part of a series, is structured in six major sections. Section 1 introduces the scope of the series in general and the nature of the European Language Equality (ELE) project which gave rise to it. Section 2 documents the status of Maltese from different perspectives: its national status; its general typology as a language, and its current usage in the digital sphere. Section 3 provides a brief, language-independent overview of Language Technology (LT) in general, covering six major application areas that are deemed to be reflective of the state-of-the-art. Section 4 then presents a highly language-dependent view on the state of LT for Maltese, starting with an examination of its special characteristics, a brief history of LT for Maltese to date, an assessment of what is currently available for Maltese in the catalogue of the European Language Grid (ELG) platform, and finally, a sketch of projects, initiatives and LT providers. Chapter 5 then attempts to provide a comparative analysis of the status of Maltese LT with respect to all other EU languages. The final Section 6 is a summary taking the form of a SWOT analysis of the status of LT for Maltese and conclusions based on what the authors perceive as serious gaps that need to be addressed urgently. The main points here are that if LT for Maltese is to progress alongside LT for neighbouring European languages, it is crucial to address not only current gaps in tools and resources, but gaps in national support for LT that recognises the important cultural, social and scientific role that it plays.

Astratt

Dan ir-rapport huwa aġġornament għar-rapport *The Maltese Language the Digital Age* (Rosner and Joachimsen, 2012b), li kien jagħmel parti mis-serje ta' white papers tal-META-NET dwar il-lingwi tal-Ewropa. Ir-rapport kurrenti, li jagħmel parti mill-istess serje, huwa maqsum f'sitt taqsimiet prinċipalli. Taqsima 1 tagħti ħarsa ġenerali lejn is-serje u tintroduċi l-proġett European Language Equality (ELE), li minnu nibtet is-serje. Taqsima 2 tiddokumenta l-istatus tal-Malti minn perspettivi differenti: ir-rwol tiegħu bħala lsien nazzjonali; il-karatteristiċi tipoloġiċi ġenerali tiegħu; u l-użu tiegħu fl-isfera diġitali bħalissa. Taqsima 3 tagħti ħarsa qasira lejn it-Teknoloġija tal-Lingwa (Language Technologies; LT) b'mod ġenerali li mhux marbut ma' lingwa partikolari, u tkopri sitt tipi ta' applikazzjonijiet li jixħtu dawl fuq l-istat kurrenti tat-teknoloġija. Taqsima 4 imbagħad toffri perspettiva fuq l-istat tal-LT fil-każ speċifiku tal-Malti. Tibda billi tgħarbel il-karatteristiċi partikolari tagħha fl-isfond ta' rakkont fil-qosor tal-istorja tal-LT għall-Malti sal-lum, b'analiżi tal-għodod disponibbli għall-Malti fil-katalogu tal-pjattaforma European Language Grid (ELG). Fl-aħħarnett, tagħti stampa ta' x'inhuma l-proġetti, inizjattivi u provvedituri fil-qasam tal-LT. Taqsima 5 mbagħad tipprova toffri analiżi komparattiva tal-istatus tal-LT għall-Malti meta mqabbel mal-lingwi l-oħra kollha tal-UE, permezz ta' indiċi speċjali, imsejjaħ Digital Language Equality index (DLE), li ġie maħluq għal dan il-għan mill-proġett ELE. It-taqsima finali, Taqsima 6, hija reċensjoni li tieħu l-forma ta' analiżi SWOT tal-istatus tal-LT għall-Malti, b'konklużjonijiet abbażi ta' dak li l-awturi jaraw bħala l-iktar nuqqasijiet serji, li jeħtieġ jiġu indirizzati b'mod urġenti. Il-konklużjoni ewlenija hi li sabiex ikun hemm progress fil-qasam tal-LT għall-Malti, pari passu mal-iżviluppi f'lingwi oħra, hemm bżonn li jiġu indirizzati mhux biss nuqqasijiet fejn jidħlu għodod u riżorsi, imma wkoll fejn jidħol is-support nazzjonali għall-LT u l-għarfien tal-importanza kulturali, soċjali u xjentifiku tiegħu.

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 European and national/regional levels, language communities, journalists, etc. and it seeks not only to delineate the current state of affairs for each of the European languages covered in this series, but also, 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 has provided a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed within the framework of the European Language Equality (ELE) project.[2] With a large and all-encompassing consortium consisting of 52 partners covering research and industry in all European countries as well as all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda and roadmap for achieving full digital language equality in Europe by 2030.

2 The Maltese Language in the Digital Age

2.1 General Facts

Maltese, locally known as il-Malti, is an official language of the EU and the national language of the Maltese archipelago comprising three islands (Malta, Gozo (Għawdex) and Comino (Kemmuna) which together cover an area of 315.6 km^2). Hereafter we will use "Malta" to refer to all three islands.

The Maltese and English languages, as well as Maltese Sign Language, are the official languages of Malta which the Administration may use for all official purposes (laws, official publications, Court proceedings etc.). Furthermore, any person may address the Administration in any of the official languages and the reply of the Administration shall be in that language.

According to a 2021 survey carried out by the National Council for the Maltese Language,[3] 97% of the Maltese population (ca. 400,000 people) consider Maltese to be their mother tongue. Although this figure signals a positive trend for the future of the language, it has to be understood as referring to informal spoken communication between adults.

The figures for written communication are somewhat different and illustrate the tendency to use English particularly for writing and also as the subject matter becomes more formal. Thus, when it comes to reading only 32% prefer to read a printed or online newspaper in Maltese, with 28% preferring to read in English. Furthermore with respect to writing formal letters or emails, the majority (54%) prefer English, with 20% opting for Maltese.

Maltese is also spoken by communities in Australia, Canada, the USA and the UK which arose after the Second World War, when large numbers of Maltese emigrated. Available statistics indicate that the total number emigrating during the period 1946–1996 to the above countries was 116,000,[4] from which one might estimate that the number of Maltese speakers currently outside Malta is around 100,000.

[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

[3] http://www.kunsilltalmalti.gov.mt/file.aspx?f=343

[4] http://www.maltamigration.com/statistics/?s=4A624EE1-7D7101215028-ACE

To give some perspective to the local figures, a Eurobarometer survey (2006) had reported that 88 percent of the population speak English “well enough to hold a conversation”,[5] 66% speak Italian, and 11% speak French. A slightly earlier study (i Capdevila, 2004) had discovered that 86% of the population express a preference for Maltese, 12 percent for English, and 2 percent for Italian.

Both Maltese and English are used within the state education system. However, the pattern of usage varies according to a wide variety of factors including the kind of school (which could be state, church, or private), and the level, subject and formality of the teaching situation. Thus within the state system, Maltese tends to dominate in primary schools, in middle school the usage is mixed, whilst at tertiary level, the University of Malta is officially an English-speaking institution (even though Maltese will often be spoken outside the lecture room). Within the school system, English is favoured for science and mathematics, whilst Maltese tends to predominate for the Art subjects, religion, and the study of Maltese itself.

Maltese is derived from late medieval Sicilian Arabic with Romance superstrata, and is often referred to as a “mixed” language due to the large number of loan words from Italian, English and to a lesser extent, French. Yet, it is primarily a Semitic language insofar as it shares underlying morphosyntactic and lexical characteristics with other Semitic languages such as Hebrew, Arabic, Amharic and others. For example, all such languages make use of root-and-template morphology whereby various forms of the same lexeme are formed by “interdigitating” vowels between a fixed sequence of root consonants. Thus, *kiteb, ktibt, kitbu* are all formed from the underlying consonant template *k-t-b*.

Maltese is the only official language of the EU that is Semitic. However, in contrast to all other Semitic languages, the Maltese alphabet is based on the Latin one with the addition of some letters with diacritic marks and digraphs (ċ, ġħ, ż, ġ, ħ). It contains 30 letters: 24 consonants and 6 vowels (a, e, i, o, u, ie).

According to Fabri (2011), the writing systems used for Maltese were somewhat ad-hoc before 1920, being based mostly on Italian, with certain additions that sometimes included Arabic symbols. In 1920, *L-Għadqa tal-Kittieba tal-Malti* (the Union of Maltese Writers) was set up with the specific aim of creating a standard orthography. Their system came to be known as *l-Alfabett tal-Għaqda tal-Kittieba tal-Malti* (the Alphabet of the Union of Maltese Writers), and, in 1924, the Government agreed to publish it. To this day, this publication is considered to be the authority on Maltese orthography. Generally speaking, a certain degree of consistency among writers and in publications became a reality in the 1950s, after the Second World War (Cauchi, 1994).

2.2 Maltese in the Digital Sphere

Characterising the Digital Sphere

It is not completely straightforward to assess the current state of the Maltese language within the digital sphere for several reasons. One is that the sphere is extremely broad in scope, encompassing a large number of different aspects of usage including, in rough chronological order: word processing, email, written and audio media, internet search, corporate websites, eServices, and of late, social media, profiling, and communication in general. All of these things are language specific and employ varying amounts of language technology, machine learning techniques etc.

Any assessment must thus make clear which of these aspects it seeks to emphasise. To take a simple example, when the digital sphere consisted mainly of word processing, one of the most severe obstacles to writing in Maltese and thus, to the creation of digital content, was the lack of suitable standards for the representation of Maltese characters and for the layout

[5] https://timesofmalta.com/articles/view/maltese-rank-with-best-language-skills-in-the-eu.76576

of a Maltese keyboard. These problems were essentially resolved when the standards were published in 2002,[6] being fully taken up by 2005. With that obstacle removed, the digital sphere of "mundane content" began to appear in larger quantities.

This is but one example of how barriers to filling out the digital sphere can be dismantled by relatively simple technical measures. With the other aspects mentioned above, however the barriers may be more complex, created by social rather than technical pressures. For example, the 2021 survey carried out by the Council for the Maltese language[7] reported that when using social media, the proportion that exclusively use Maltese is 33.8% which is significantly lower than the proportion that consider Maltese their native language.

A second reason why the penetration of Maltese within the digital sphere is hard to assess is that the distinction between local internet users, and local users *of the Maltese language* on the internet is rarely reported upon. One exception is the study by Cortis and Davis (2021) who reported the results in Table 1 concerning online comments to the annual budget. The high figure for English is expected given that the media concerned were mainly English speaking. Of more interest here are the very low figures for pure Maltese, and the much higher figure for code-switched English/Maltese

Language	Budget 2018	Budget 2019	Budget 2020
English	71.52	71.55	79.99
Maltese	4.34	6.22	3.21
Codeswitched	23.2	21.24	15.97
Other	0.93	0.99	0.83

Table 1: Percentage distribution of language annotations for online comments

Certain other more general facts can also be stated. For example, according to the National Statistics Office[8] internet usage in Malta is now (2020) at par with other EU member states, standing at 86.9% of the population. This figure has steadily increased since 2011 when it was only 66%.

Online News Media

Online content in Maltese certainly exists in considerable quantity. There have always been several Maltese newspapers (e. g. *KullHadd, L-Orizzont, In-Nazzjon, Illum*). In addition, the broadcast media (radio and TV) are almost exclusively in Maltese. Since the previous language report, the main developments can be summarised thus: (i) there has been a general decline in hard-copy newspaper readership; (ii) all the media are now available online and the majority of readers prefer the online version (iii) various online-only news websites have appeared, one of which (Newsbook[9]) operates in both Maltese and English; (iv) the full Maltese character set is now used by the vast majority of Maltese news media.

Social Media

The social media are widely used (97% of the population according to a 2021 survey by Datareportal.[10]) Misco, a leading provider of information to the business community, reports the usage pattern shown in Table 2.

[6] MSA (Maltese Standards Authority) 100:2002 Maltese Keyboard Standard

[7] http://www.kunsilltalmalti.gov.mt/file.aspx?f=343

[8] https://nso.gov.mt/en/News_Releases/Documents/2021/02/News2021_028.pdf

[9] http://newsbook.com.mt

[10] https://datareportal.com/reports/digital-2021-malta

Social Media Network	Users (% of population)
Facebook	98
Messenger	85
Linkedin	66
Instagram	62
pinterest	24
Twitter	13

Table 2: Social media network usage in Malta

Facebook remains the most accessed social media network, but there is a trend of increased usage of Instagram and YouTube. In contrast to many other EU countries, Twitter usage in Malta is remarkably low. Increasingly, usage of social media typically takes place on a smartphone or a laptop and less on a desktop PC or a tablet.

There is also a gap between social media creators and non-creators: While nearly nine of ten of people in Malta go online at least once a day, according to National Statistics Office data, 64% of social media users only view other persons' content and comments, without taking further action such as sharing or creating new content. Meanwhile, the use of certain apps, such as social networking apps, news apps and retail apps, is increasing, while the usage of others, such as travel apps, is decreasing, falling sharply in 2020 (no doubt as a result of the pandemic).

Content Creation: Wikipedia and Youtube

These trends are to some extent reflected in the Maltese Wikipedia statistics.[11] The Maltese Wikipedia currently ranks at 204/325 (for comparison, English, Portuguese, Irish, Icelandic, Romansch rank at 1, 18, 93, 95, 213 respectively). It contains nearly 4M words distributed over 4,400 content pages[12] (cf. 6.5M articles for English). This compares to about 3,000 pages in 2011 – not a huge level of growth. There are ca. 19,000 registered users. The number of active users (making changes every 30 days or less) is much smaller at ca. 40.

Another social media channel which gives rise to localised content in many other countries is Youtube. In June 2018, YouTube announced the launch of a country-specific version and local domain for Malta (youtube.com.mt). This initiative was followed in March 2019 by YouTube's announcement that its YouTube Partnership Programme (YPP), which caters for the monetisation of content was to be made available to the ecosystem of local content creators in Malta.[13] The Maltese government announced discussions on how the partnership programme could benefit Maltese artists and content creators at par with other European citizens. However, to date the country-specific website still operates in English rather than Maltese and although some content in Maltese exists, the volume is limited.

Clearly, a fundamental problem here is that contributors tend to be remunerated in proportion to the size of their audiences, and Maltese-speaking audiences are small by definition. Hence, it is not surprising that content producers tend to increase outreach by using English rather than Maltese.

[11] https://meta.wikimedia.org/wiki/List_of_Wikipedias#All_Wikipedias_ordered_by_number_of_articles

[12] this is about 25% of the total number of pages in the wiki which also include discussions, referrals etc.

[13] https://timesofmalta.com/articles/view/youtube-partnership-programme-now-available-in-malta.704787

Kelma Kelma

A renowned online page which has successfully bucked this trend is Kelma Kelma[14] which started in January 2013 as a Facebook page run by Dr Michael Spagnol, now Head of the University's Department of Maltese. It gathers many interesting aspects of the Maltese language, such as the word of the day, proverbs, idioms, word games, and presents them in a fresh and intriguing way with colorful pictures. Over time, having captured the imagination of the nation, the page emerged from the world of social media and continued to promote Maltese in schools, on radio and television, in newspapers and magazines, and on stage. It has won several national prizes. In 2018, on its fifth anniversary, the website kelmakelma.com was launched which is a collection of information on vocabulary, grammar, history, and many other curiosities that shed light on the beauty of the Maltese language.

The .mt Domain

.mt, the top-level country domain for Malta is administered by the Malta Internet Foundation also known as NIC(Malta) whose responsibilities are (i) creation and implementation of a domain naming policy for .mt; (ii) maintenance of the .mt nameservers; (iii) promotion of the .mt domain; (iv) registration of new domain names.

There are currently ca. 17,000 domain names registered under .mt and its subdomains (.com.mt, .edu.mt, .org.mt, gov.mt), more than three times the figure in 2010. This can partly be attributed to changes in pricing and opening up of the top-level (.mt) for registration, which was previously restricted to the second-level domains listed above.

3 What is Language Technology?

Natural language[15] 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, mathematics, psychology and notably, AI, amongst 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 a starting point in the 1950s alongside Turing´s renowned writings on computing machinery and intelligence (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

[14] http://kelmakelma.com

[15] 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).

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 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). A very popular application 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 just very 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.[16]

4 Language Technology for Maltese

Intrinsic Characteristics of Maltese

As mentioned earlier, Maltese is the only official Semitic language in the European Union and the only Semitic language written in a Latin alphabet. The Maltese alphabet makes use of some special graphemes that differ from other Latin alphabets (the sound values are given in the International Phonetic Alphabet): ċ tʃ, ġ dʒ, għ (mostly silent), ħ h, ż z. (Fabri, 2011; Borg and Azzopardi-Alexander, 1997). Some particular characteristics of Maltese are:

- Free word order

  Even though there are no case endings, Maltese has a very free word order. The sentence *Il-kelb gidem il-qattusa lbieraħ* ('The dog bit the cat yesterday.') has the word order S(ubject) V(erb) O(bject) but could also be expressed with orders VOS (*Gidem il-qattusa l-kelb*) and OVS (*Il-qattusa ngidmet mill-kelb*). The different word orders are all acceptable but emphasise different aspects of the meaning. For example, OVS emphasises the object for contrast.

- Mixed morphology

  One of the unusual features of Maltese is the mixture of stem-based and Semitic (root-and-pattern-based) morphology. Stem-based morphology forms words by concatenating affixes to a stem, as seen in many Romance languages (e. g. *-i* for plural). For the Semitic component, the basic "unit" within a word is not a stem but a root made up of three (sometimes four) consonants in a fixed order that carry a general meaning. Word stems with their specific meaning are formed by arranging the consonants according to a certain pattern. For example, the root *k-t-b* carries the meaning of everything connected with "writing" and the pattern **1v2v3** (where the numbers represent the root consonants and each **v** a vowel) can be 'applied' to *k-t-b* and *i-e*, to yield the perfective verb *kiteb* 'he wrote'. Inflection of this verb for person and number takes place by affixation e. g. the plural affix *-u*, giving the form *kitbu* 'they wrote'. Different patterns can produce different forms. Thus, the pattern **1v22v:3** (**v:** stands for a long vowel) to the root renders the agent noun *kittieb* 'writer'.

[16] 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).

The plural in Maltese can also be formed both by affixation (e. g. *student* – student; *studenti* – students) and Semitically (the so-called broken plural forms), i. e., no affixation takes place, but the noun is changed internally, e. g. *ktieb* – book vs. *kotba* = books. Some words even have both forms of the plural *tapit* – carpet; *tapiti, twapet* – carpets.

Large numbers of loan verbs are imported using a special verb class that can accommodate undigested stems Mifsud (1995). For example, the English stem *park-* became the basis of the Maltese verb forms *pparkjajt, pparkjat, pparkja* 'I/you, she/ he parked'. Today, this formerly marginal Semitic special verb class has increased in size due to the influx of English loan verbs. It is highly productive, often giving way to *ad-hoc* loans of English verbs even when they already have a Semitic counterpart in Maltese. For example 'to download (a file)' can be perfectly expressed using the Semitic verb *niżżel* (originally meaning 'he caused to come down'). However, taking the English stem *download* and importing it via the special verb class instead gives forms like *ddawnlowdjajt, ddawnlowdjat, ddawnlowdja* 'I/ she/ he down-loaded'. This strategy is often criticised as corrupting the language. Fabri (2011).

- Aspect-based temporal system

  Verbs in Maltese are marked for aspect, i. e., as to whether an action is completed (perfective) or not completed (imperfective) – for a full account of tense and aspect in Maltese, see Fabri (1995) and Ebert (2000). In the absence of any other grammatical markers, verbs in the perfective are interpreted as "past tense" and verbs in the imperfective as "present tense": *Andrew kiteb* 'Andrew wrote'; *Andrew jikteb* 'Andrew writes'. Combination of the imperfective verb with *kien*, the perfective form of the verb for 'to be', expresses habitual past: *Andrew kien jikteb* 'Andrew used to write'. Adding word *qed* 'progressive' (like the English *-ing* form) gives *Andrew kien qed jikteb* 'Andrew was writing' etc.

- Lack of a morphological infinitive

  Maltese verbs do not have morphological infinitives. Thus, in complex predicates like in the English sentence 'Andrew wants to write', both verbs are morphologically finite: *Andrew irid jikteb* (literally: 'Andrew he wants he writes') even though semantically, *jikteb* is not finite.

4.1 Brief History

Period 1997-2010

Table 3 summarises the main enablers and contributions to Maltese Language Technology up to ca. 2010. Most of these were at the level of University research prototypes in the subfields listed. An EU research project (LT4eL) focused on the semantic indexing of learning materials. Further details of these efforts are provided in the previous META-NET language report for Maltese (Rosner and Joachimsen, 2012a).

The last entry in the table refers to the LREC2010 conference which represented something of a watershed moment for LT in Malta. For the first time, a large, prestigious LT conference was being hosted on the island, bringing an unprecedented number of researchers together. The event did not pass unnoticed, particularly by the office of the President and also by the international community.

At the end of this period, a written corpus of around 1.5 million tokens had been made available through the MLRS resource server, which also offered simple resource-submission facilities to contributors, and some basic text-processing services such as a tokeniser. Work

Year	LT Subfield	Reference
1997	Speech Synthesis	Micallef (1997) Phd Thesis
1998	Computational Lexicography (Maltilex)	Rosner et al. (1998)
1999	Rule-Based Morpological Analysis	Mangion (1999)
2000	Rule-Based Machine Translation	Farrugia (2000)
2001	Spelling Correction	Mizzi (2000)
2004	Statistical Machine Translation	Bajada (2004)
2005-8	LT4eL Project (EU Cordis)	Vertan et al. (2007)
2008	Maltese Language Resource Server Version 1	Rosner (2008)
2009	Lexical Information Extraction	Camilleri and Rosner (2009)
2010	TTS Speech Synthesis (ERDF Project)	Borg et al. (2014)
2010	LREC 2010, Valletta	http://www.lrec-conf.org/lrec2010/

Table 3: Summary of Contributions to LT for Maltese 1997-2010

had begun on the Maltese Speech Engine (MSE), a TTS system for Maltese supported by ERDF funds (2009-2012).[17]

However, despite these successes, there were many shortcomings. The corpus was comparatively small and not fully representative, being predominantly textual and monolingual, and lacking genres such as academic text and works of fiction. Development of the analysis tools required to form a basic text processing pipeline for content extraction were lacking, hampered *inter alia* by lack of agreement about a tagset, and insufficient annotated training materials. In contrast to the situation for speech synthesis, automated speech recognition was neglected. Finally, besides the research prototypes mentioned above, there was no serious appetite, nor funds, for further work on Machine Translation.

Period 2011-2016

Whilst LREC2010 was under way, the Multilingual Europe Technology Alliance (META) was being put together by META-NET, an EC Network of Excellence European Commission funded through the ICT PSP Programme. META-NET succeeded in setting up three regional subprojects, whose central objectives were the assessment and collection of datasets and software tools for speech and language processing, and their distribution and dissemination on a pan-European digital platform. Malta joined the consortium for the subproject METANET4u[18] which obtained funding under the ICT/PSP programme. The resources and tools for Maltese collected there were destined for META-SHARE,[19] an infrastructure that made available quality LRs and related metadata to all its members and users.

In 2012, a first public release of the the MSE speech synthesiser was delivered to The Foundation for IT Awareness (FITA) in 2012, where it continues to serve as an enabling technology for people with a disability. At about the same time, Gatt and colleagues at the University's Institute of Linguistics began revamping the MLRS resource server with a view to building corpora and related tools for Maltese on a much larger scale (Gatt and Čéplö, 2013). The main innovations were semi-automated data-collection by focused web-crawling followed by a post-processing pipeline that carried out paragraph and sentence splitting and delivered a series of tokens POS-tagged with the (now agreed) 2-level MLRS tagset that included 41 major categories. The tagger itself was trained on ca. 28K words using TnT (Brants, 2000) and achieved 95% accuracy. Some effort was also made to balance the corpus using donated text from authors and users. The resulting Maltese Corpus, known as Korpus Malti v2.0, was

[17] https://fitamalta.eu/projects/maltese-speech-engine-synthesis-erdf-114/
[18] https://www.di.fc.ul.pt/~ahb/metanet4u/index.html
[19] http://www.meta-share.org/p/66/About

released in 2013 and contained ca. 130 million tokens, in various genres, annotated with PoS tags, lemmas and morphological roots. Being implemented using the IMS Corpus Workbench,[20] it also included some sophisticated pattern-based search facilities.

On the syntax-semantics frontier, a usable computational grammar of Maltese[21] was incorporated into the Grammatical Framework's[22] Resource Grammar Library (Camilleri, 2013), thus opening Maltese to many multilingual applications that have been developed using using GF.

A third version of the Maltese Corpus (Korpus Malti v3.0) was released in 2016, substantially larger (ca. 250 million tokens) and somewhat more representative than its predecessor, and tagged with the Maltese Tagset v3.0, developed by Slavomir Céplö and Albert Gatt which includes annotations for lemma and root. MLRS also includes other corpora, notably

- CLEM, a one million token Corpus of Learner English in Malta, consisting of English essays by students. The corpus is stratified by gender, school type, candidate's region of residence, date of birth and mark/grade. Tokens are annotated with POS, lemma and orthographic errors.

- Ġabra: An Open Lexicon for Maltese: A Maltese-English full-form lexicon, grouping entries by lemma and morphological root, with English glosses and inflectional forms. Ġabra also provides a web service API for querying and lemmatisation, for developers. The database is downloadable. Ġabra was originally built by John J. Camilleri as an opportunistic collection, partly based on automatically generated wordforms and partly on other resources.

- Dizzjunarju tal-Malti[23]: Ġabra, underwent substantial redevelopment as part of this project, as a result of collaboration between Malta Communications Authority, the Institute of Linguistics, the Department of Intelligent Computer Systems at UM, as well as the National Council for the Maltese also available in a mobile-friendly version and as a downloadable application from Google Play to allow use of this service anytime, any place.

- The Dictionary of Maltese Sign Language Language.[24]

2015 marked the beginning of Malta's involvement with ELRC. The first country workshop took place in February 2016 with a relatively small audience of 37 participants.

Also in 2016, the IT subcommittee of the Council for the Maltese Language compiled a report and roadmap for the development of Digital Language Resources and Tools for the Languages of Malta[25] whose three main recommendations were

1. The creation of a central repository of language resources and tools related to Maltese, as well as the other languages used in the Maltese islands, notably English and Maltese Sign Language;

2. The setting up of an initiative, overseen by the National Council for the Maltese Language, to bring together stakeholders and ensure the long-term curation of language resources and tools in the Maltese context;

3. The involvement of more stakeholders and the sensitisation of the public as to the availability and importance of such resources.

[20] https://cwb.sourceforge.io/cpqweb.php
[21] https://www.giters.com/johnjcamilleri/Maltese-GF-Resource-Grammar
[22] http://www.grammaticalframework.org
[23] http://www.maltesedictionary.org.mt
[24] https://mlrs.research.um.edu.mt/resources/lsm
[25] http://www.kunsilltalmalti.gov.mt/file.aspx?f=309

Although some effort has been made towards putting these recommendations into practice, they have not been fully realised. MLRS to some extent addresses the first issue, but it remains largely the work of volunteers, and is in no way sanctioned officially. Regarding the second and third points, the community of LT users/contributors remains largely fragmented and so do their outputs.

Period 2017-present

This period is marked by incremental steps towards the development of a dependency parser, and membership in various EU projects and initiatives pertinent to the further development of LT for Maltese, as follows:

- **Dependency Parsing.** The creation of a Universal Dependency Treebank for Maltese (Čéplö, 2018) consists of just over 2000 sentences annotated with dependency parse trees following the UD annotation guidelines.[26] This dataset permitted a series of computational experiments Zammit et al. (2019) to be carried out leading to the expected release of a prototype dependency parser for Maltese in 2022.
- **Multiword Expressions.** Malta participated in COST action IC1207 (2013-1017) PARSEME whose aim was to improve linguistic representativeness, precision and computational efficiency of NLP applications by studying Multi-Word Expressions (MWEs). UM participated actively in the writing of a state-of-the-art survey (Constant et al., 2017).
- **Language and Vision.** Malta participated in COST action IC1307 (2014-2018) – The European Network on Integrating Vision and Language, resulting in the setting up of the RIVAL research group[27] which developed Face2Text, a prototype system and annotated dataset for generating verbal descriptions of facial images (Gatt et al., 2018).
- **ELRC Workshops**. Workshop 2 took place in 2019, and focused on live demonstration of the newly released neural eTranslation. Workshop 3 (2022) emphasised the practical utility of eTranslation and, succeeded in drawing a notably large online audience from the public administration.

Other recent initiatives and projects and programmes are discussed in Section 4.3.

4.2 Language Data and Tools

This section aims to give an overview of resources that are currently available in the ELG catalogue. For the sake of feasibility, we have been very selective in our approach. In general, we follow the classification used by ELG, and for each of these, have chosen some representative examples to illustrate some salient characteristics. Where there is a need to refer to individual resources, we use the ELG name which can be searched and found directly from within ELG.

The total number of language resources available for Maltese is approximately 200: a very small number compared to languages for other countries with larger numbers of speakers, as shown for selected countries in Figure 1.

Nevertheless, such statistics can be misleading when adjusted for the number of speakers, as shown by the very different ranking shown in Figure 2.

Of these, text corpora dominate, being roughly twice as numerous as tools or lexical resources as shown clearly in Figure 3.

[26] https://universaldependencies.org/guidelines.html

[27] https://rival.research.um.edu.mt

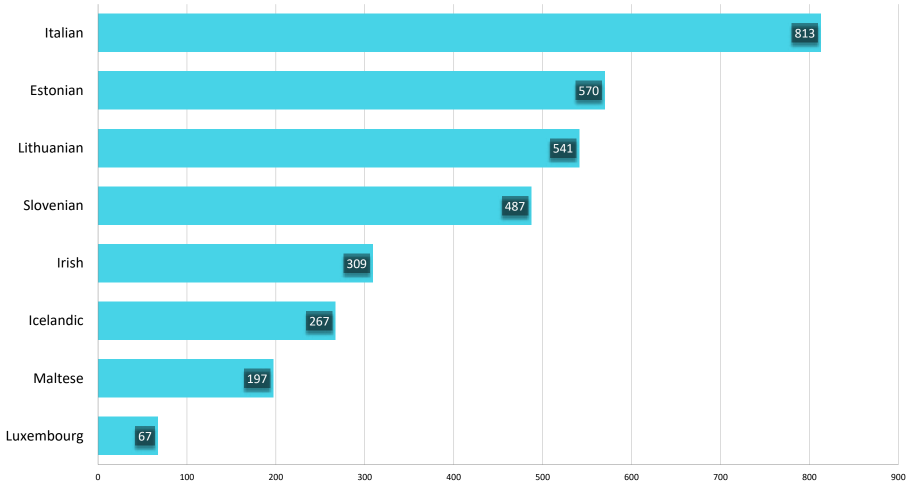

Figure 1: Number of available resources in selected countries

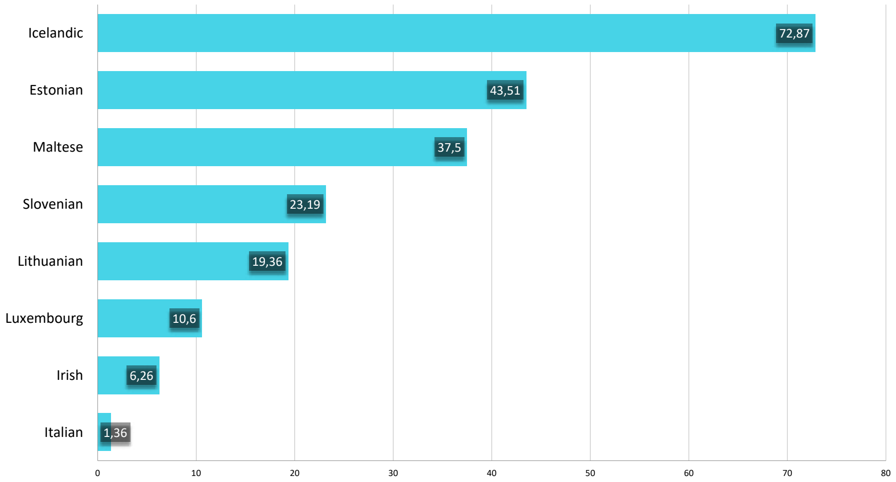

Figure 2: Number of resources per 100K population

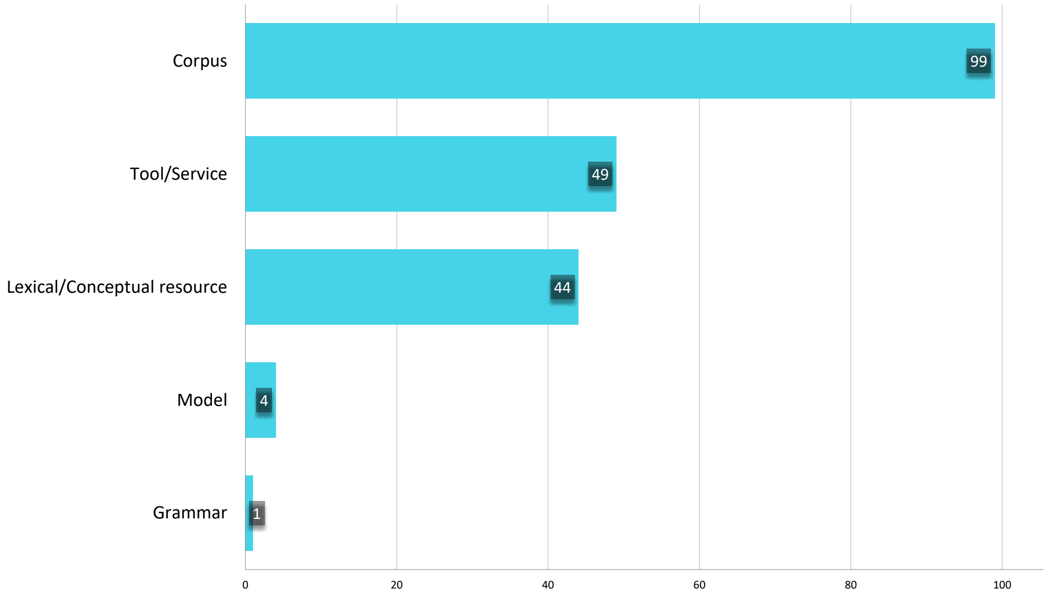

Figure 3: Number of Maltese Resources in ELG by type

Corpora

There are approximately 100 corpora for Maltese available though ELG of which the vast majority are written text. A few are spoken, and yet fewer are multimodal.

Monolingual Corpora

Many of the monolingual corpora listed below actually form part of *multilingual* collections. There are several unannotated, essentially raw collections of running text such as Maltese Wikipedia dump, Open Super-large Crawled Aggregated coRpus – Maltese from the OSCAR project. Another example is Sketch Engine Maltese Corpus mtWaC with a total size of 110 million words.

Corpora in this category are mostly by-products of projects and annotated for training with respect to some particular task e. g. MWE identification (Annotated corpora of the PARSEME Shared Task) or POS Tagging (MLRS POS Gold data using the XPOS tagset), anonymisation (with fine-grained annotations of anonymisable segments from the recently completed MAPA project), morphological analysis (UniMorph – Maltese containing morphological annotation of Wikipedia text – part of the UniMorph project (Kirov et al., 2018) project aiming to annotate morphological data in a universal schema, NER (WikiAnn – Maltese (Pan et al., 2017) with silver annotations for automated NER tasks, and language identification (wili_2018, a benchmark dataset for language identification and contains 235,000 paragraphs of 235 languages. Finally we have the Maltese Universal Dependencies Treebank which contains manual annotations of POS (UPOS & XPOS) for Dependency Parsing. This is part of the multilingual Universal Dependencies corpus

Bilingual and Multilingual Corpora.

Many Maltese text corpora in this category have been created as a result of activity at EU level, prime examples being the JRC-Acquis Multilingual Parallel Corpus, which comprises the total body of EU law applicable in Member States. and, also in the domain of law, the Parallel corpus collected from the European Constitution.

Other notable bilingual and multilingual collections are: (i) COVID-19-related corpora compiled by different EU agencies (e. g. European Medicines Agency (EMA), European Parliament (EP)) dealing most prominently with health, legalistic, and Parliamentary issues. All are released with Creative Commons Attribution 4.0 International licences; (ii) Paracrawl (EU CEF) 11 bilingual parallel corpora aimed at broader/continued Web-Scale provision of Parallel Corpora for European Languages; (iii) bilingual corpora for all EU Languages for training of NTEU Machine Translation engines, 9 of which involve Maltese (Bié et al., 2020); (iv) Tilde MODEL Corpus – Multilingual Open Data for European Languages(Rozis and Skadiņš, 2017) collected from sites allowing free use and reuse of content, and Public Sector web sites.

There are also sentence-aligned corpora, some of which (DGT-Translation Memory, DG-EAC Translation Memory, from the Directorate General for Education and Culture and the European Union European Centre for Disease Prevention and Control ECDC respectively) take the form of translation memories represented in TMX format.

National corpora include bilingual English/Maltese Government Corpora e. g. Laws of Malta, Government Gazette.

Finally we mention evolving resources which currently contains too few sentence pairs involving Maltese to be useful e. g. tatoeba (215 sentence pairs); ted_talks_iwslt (bilingual captions for one Ted talk).

Speech Corpora

The two speech corpora listed as such are (i) CommonLanguage, composed of 1 hr of open access speech recordings for each of ca. 40 different languages, and (ii) CommonVoice (Ardila et al., 2020), Mozilla's initiative which also includes transcriptions and employs crowdsourcing for the collection and validation of data. All of the speech data is released under a Creative Commons CC0 license, There are also two recent corpora from the MASRI project (spontaneous speech and read speech) in Maltese with transcriptions. However, these are not listed as speech resources.

Lexical and Conceptual resources

Several resources of the resources in ELG are lexical but nevertheless listed as corpora. These include (i) BabelNet, a multilingual encyclopedic dictionary. BabelNet 5.0 covers 500 languages and is obtained from the automatic integration of several resources such as WordNet, Wikipedia, Wiktionary, Wikidata, GeoNames; (ii) terminological databases such as ms_terms used to develop localized versions of applications that integrate with Microsoft products and to integrate Microsoft terminology into other terminology collections (iii) senti_lex: sentiment lexicons generated via graph propagation based on a knowledge graph

Models and Grammars

Universal Dependencies 2.4 and 2.5 Models for UDPipe Tokenizer, POS Tagger, Lemmatizer and Parser models.

Machine Translation

19 Machine translation tools are available which are (i) commercial (with free use): BING, Google Translate, Collins Translator, Prompsit Translator (alternative to Apertium); (ii) European: CEF e-Translation which includes both web and API interfaces; (iii) Enterprise-grade tools such as ModernMT: based on Fairseq Transformer model; Enterprise cloud solution for professional translators, accessed via API or plugin; SDL Machine Translation. Open Source edition available. SYSTRAN (also with free online service); (iii) ELG-compatible Translation Services: HelsinkiNLP – OPUS-MT; finally (iv) Aggregators which provide an interface to existing translation solutions: Qtranslate (free translator for Windows. with other services such as image text recognition, Text-to-speech, semantic search.

NLP Pipeline Services and Tools

Under this rubric we include the basic machinery needed to perform analysis starting with raw text e. g. paragraph and sentence splitting, tokenisation; POS-tagging; Named Entity recognition: MLRS services; LIMA – Libre Multilingual Analyzer; STANZA; UDPipe Maltese; MAPA NER for Maltese; EvidenSSE; Text Tonsorium.

Some of these include multi-word token expansion, lemmatisation, part-of-speech and morpho-syntactic feature tagging, and semantic search.

Accessibility

Most data resources and services mentioned above are freely available, either via a web interface, by means of a downloadable tool, or through an API. Many permit direct download, whilst others provide an email contact through which the resource may be obtained. Many do not mention licensing conditions at all, but, as shown in Figure 4, those that do use well-known open source licences, the most popular is CC-BY-4.0 (49 resources).

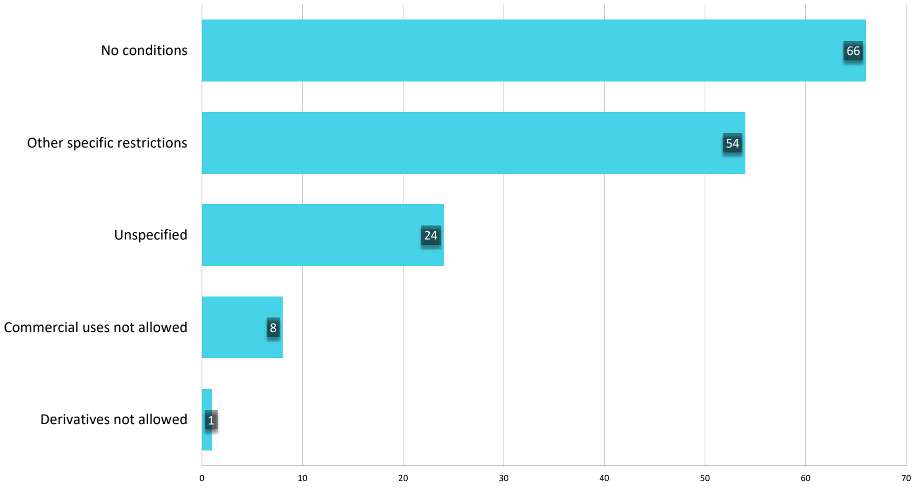

Figure 4: Number of Maltese Resources in ELG by licence type

4.3 Projects, Initiatives, Stakeholders

Current National Initiatives

There are a number of projects and initiatives currently being undertaken in the efforts to develop the necessary tools and resources to computationally process Maltese.

- **National Programme for AI.** In October 2019, the Maltese Government launched the National AI Strategy for Malta,[28] with the intention to gain strategic economic and competitive advantage in the field of AI. One of the main pillars of the strategy is to create an AI ecosystem infrastructure based on investment and innovation. At the centre of this pillar, the strategy commits to invest in tools to enable Maltese Language AI solutions, with funds committed to the development of Maltese language technology solutions.
- **MDIA Agreement.** The Malta Digital Innovation Authority (MDIA[29]) entered into an agreement with UM so that the necessary Maltese language technology tools can be developed. These tools will focus on morphological analysis, dependency parsing, named entity recognition and part-of-speech tagging. The undergoing project is currently investigating the performance of the latest type of neural approaches and exploring whether a single monolingual neural model can be developed and then fine-tuned to the respective tasks. Another avenue being researched is the influence of multilingual models that were not trained on Maltese and the extent of their impact on the processing of Maltese. This work is expected to be published by the end of November 2022.
- **KMM**. The MDIA agreement also dedicated funds to the development of the first Spoken Maltese Corpus (Korpus Malti Mitkellem – KMM). This spoken corpus is being systematically collected to ensure that a well-balanced and well-represented corpus is made available. Again, this project is underway and results will be published in 2022.
- **Research at UM.** There are small research initiatives within UM exploring how orthographic and grammar correction models can be trained in low resource scenarios. The research is looking at ways of leveraging multilingual data and then applying neural models to Maltese. Initial results can be expected in June 2022.
- **Spellchecker.** In November 2019, the Government of Malta announced that it would be investing in the development of a spell checker for the Maltese Language. The cost was included as part of the Budget 2020.[30] Currently, there is no publicly available information with respect to the progress of this initiative. However, the Government of Malta continues to place strong emphasis on the development of digital tools for Maltese. The development of a spell-checker for the Maltese language will be crucial in the continued use of Maltese in the digital sphere, especially in the business domain where most communication is carried out in English.

Recently Completed National Projects

- **MASRI.**[31] With a view to developing ASR for Maltese, the project has created speech corpora and investigated data augmentation techniques, and neural methods for speech-to-text production. The project has delivered MASRI-HEADSET a fully annotated speech corpus Hernandez Mena et al. (2020) and developed a grapheme to phoneme tool for

[28] https://malta.ai/wp-content/uploads/2019/11/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf
[29] https://mdia.gov.mt
[30] https://timesofmalta.com/articles/view/budget-2020-maltese-language-spell-checker-and-drinking-fountains-in.742092
[31] https://www.um.edu.mt/projects/masri/

Maltese which is available on the MLRS server. The project was supported by a UM Research Fund Excellence grant.

- **Dictionary of Maltese Sign Language**.[32] The online Maltese Sign Language (LSM) dictionary is intended as an important resource for the Deaf community and all who are learning the language. The dictionary contains entries searchable through Maltese and English glosses. These glosses are the closest equivalent label in English (or Maltese) to the meaning of the sign in LSM.
- **MAMCO**.[33] MAMCO (Paggio et al., 2018) is a multimodal resource involving Maltese conversational data which explores the interaction between speech and gesture in first encounter conversations.

Current EU Projects and Programmes

- **LT-Bridge**.[34] H2020 Widespread project 2021-2023 to integrate UM's Department of Artificial Intelligence (DAI) and Institute of Linguistics and Language Technology (ILLT) into the European research community in the area of AI-based language technologies.
- **LCT**.[35] Erasmus Mundus Master Programme in Language and Communication Technologies 2019-2025). 2 year Masters studying one year each at two different European universities of the consortium. Students obtain two Master of Science/Arts degrees approved in the respective countries of issue.
- **ELRC**.[36] Manages, maintains and coordinates the relevant language resources in all official languages of the EU and CEF associated countries.
- **Nexus Linguarum**.[37] COST action to promote synergies between linguists, computer scientists, terminologists, and other stakeholders in industry and society, to investigate and extend the area of linguistic data science.
- **NLTP**. The main aim of the National Language Technology Platform (NLTP), funded under CEF, is to build a National Language Platform for Maltese which will integrate the eTranslation services developed by the European Parliament and fine-tune the translation memories using more local parallel data. Apart from building custom translation memories for the local public administration sector, the platform will also be used as a point for centralising all the different language technology services being developed. The main Maltese partner in the NLTP project is the Malta Information Technology Agency (MITA), a government IT agency which is tasked with developing and implementing the necessary infrastructure to enhance the public service technologies. MITA is also joined by the UM, whose role is to research and develop custom translation memories, and the Office of the State Advocate, which has a crucial role in liaising with public entities for data provision and technology uptake. The results of the project will be made available in March 2023.

[32] https://mlrs.research.um.edu.mt/resources/lsm
[33] https://sites.google.com/view/mamcocorpus/home
[34] https://lt-bridge.eu
[35] https://lct-master.org
[36] https://www.lr-coordination.eu
[37] https://nexuslinguarum.eu/the-action

Recent EU Projects

- **MAPA**. UM participated in MAPA (Multilingual Anonymisation toolkit for Public Administrations), a European-funded project under the Connecting Europe Facility (CEF) with the project coming to an end in December 2021. The MAPA project aimed at developing an open-source de-identification toolkit for all official languages within the European Union. The toolkit relies on Named-Entity Recognition and Classification using the latest neural techniques. The University of Malta's participation in the project was centred around the contribution of annotated data for named-entities in Maltese. This was the first large-scale Named-Entity annotation effort carried out for Maltese. Moreover, since the focus of the toolkit is de-identification and anonymisation, the entity annotation was carried out at a more fine-grained detail.

National Research Infrastructures for Language / LT

There are no national research infrastructures for LT as such. However, two entities having functions that overlap those of such infrastructures are:

- The **National Council for the Maltese Language**, and in particular its IT subcommittee, whose remit could potentially influence the direction of LT research locally, as stated in the recommendations of its 2016 report.
- **DARIAH**.[38] Malta is now a member of The Digital Research Infrastructure for the Arts and Humanities which aims to enhance and support digitally-enabled research and teaching across the arts and humanities. DARIAH is a network of people, expertise, information, knowledge, content, methods, tools and technologies from its member countries. However, national activities under DARIAH depend on national funding which to date has not been forthcoming,

LT Providers

Although the use of online services which make use of LT technologies is on a par with other European countries, the number of technical LT providers is very low. In some ways, this mirrors the situation mentioned earlier concerning online content in Maltese, where the number of producers is a very small fraction of the number of consumers.

5 Cross-Language Comparison

The LT field[39] 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 unforeseeable 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).

[38] https://www.dariah.eu

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

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[40] 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.)

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

[40] 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.

4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[41]

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[42] 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.[43]

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 4 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

[41] 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.

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

[43] 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.

			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 4: 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)

national or regional level in at least one European country and other minority and lesser spoken languages,[44] 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 5 visualises our findings.

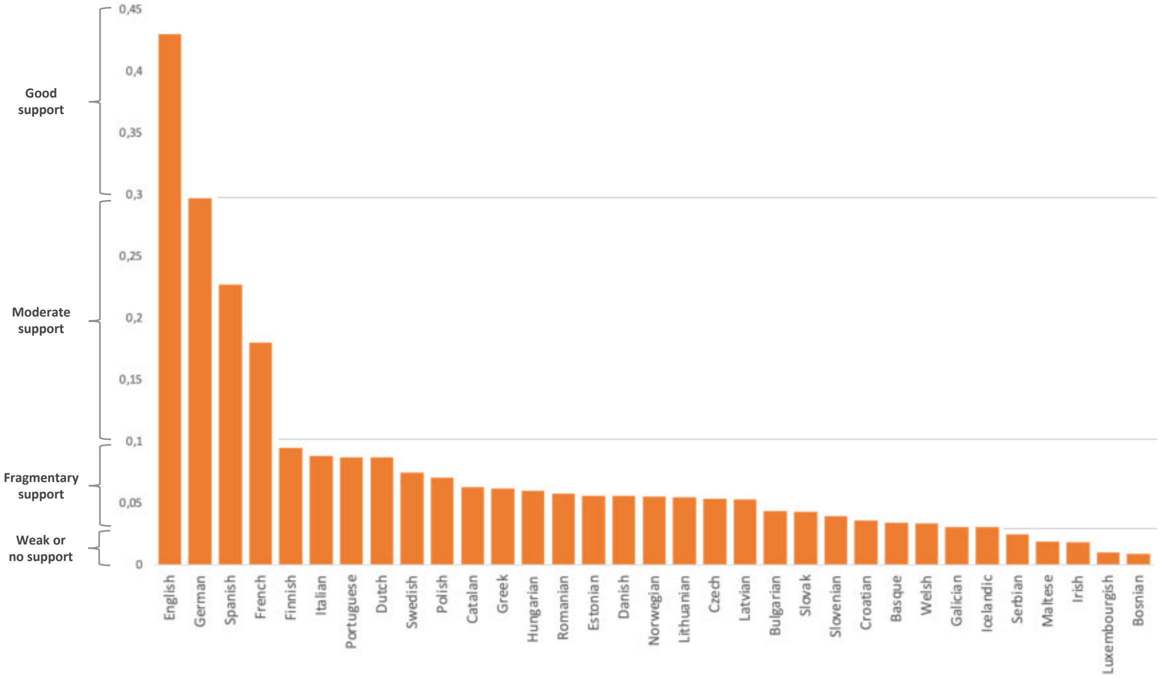

Figure 5: Overall state of technology support for selected European languages (2022)

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 in-

[44] In addition to the languages listed in Table 4, 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.

stead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e. g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

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

Maltese remains a language with a number of special characteristics of which the most predominant are the small number of speakers and the relatively constrained geographical areas in which they operate. A second major characteristic is that although Maltese is the only national language, it operates alongside English as a second official language. These characteristics have tended to dampen enthusiasm for LT as a means of lowering language barriers, since many Maltese don’t see the barriers amongst their native-language speaking selves, nor with outsiders, where English has served them well for centuries.

Perhaps these characteristics explain a comparatively low appetite for public and private investment in LT which we have witnessed up to now, which result in Malta’s relatively low position in the rankings for technology support shown in Figure 5.[45]

There is reason to believe that this situation is slowly changing as projects like NLTP begin to bear fruit, as language-sensitive access to international markets becomes the accepted norm, and as online technologies become more accessible, and more necessary, to ordinary citizens.

Strengths

Amongst the strength we must count the availability of local expertise, in technical aspects of LT and linguistic aspects of the Maltese language. We also note that access to LT-related technologies (e. g. speech synthesis, machine translation, search) has improved considerably over the last 10 years. For example, eTranslation for Maltese has become good enough to have a positive impact on anybody involved with the creation of content in Maltese. The National AI strategy has led to some commitment by Government on the substantial challenge of creating bilingual access to Government services. Finally, the imminent availability of various LT-related services for Maltese e. g. anonymisation and Named Entity recognition, mostly through EU initiatives, will eventually have a positive impact on the availability of language data resources in sensitive areas. However, this will not happen on its own. Some stimuli are required to bootstrap the integration of LT into the socio-economic framework.

[45] We should note that Table 4 reflects the current contents of the ELG catalogue and does not include certain existent resources that do not yet form part of that catalogue.

Weaknesses

Nearly all the expertise in LT for Maltese is currently concentrated at the University of Malta (and a handful of Universities abroad). Although, as mentioned above, this is a strength within that institution, it is a weakness at national level because there is not enough LT expertise in other sectors, most notably in Government and in industry, reducing the synergistic potential of LT. Hence, compared to other countries, there is a serious lack of takeup and development of LTs by Government and by local industry.

There is as yet no coherent strategy for the management and curation of resources and tools at national level. This results in a lack of continuity, and the scarcity of domain-specific corpora. There are still major problems in getting the necessary authorisations to make use of language data in certain domains e. g. health and commerce, both of which are seen as sensitive.

Opportunities

There is great potential for Maltese language services to be harnessed throughout the economy in both private and public sectors and across different domains using different modalities.

This potential of LT within industry has long been recognised by associations such as LT-Innovate (Language Technology Industry Association) who claim that future take-up will be driven from three different angles:[46] (i) language neutrality, the idea that something like a "translate" layer needs to be built into the technology stack that underpins the economic network; (ii) data markets, where language data are exploited for their inherent "languaginess," not as time series or for their numerical content, and (iii) multimodal fusion whereby different modalities of data – e. g. text, speech and image – can entangle in complex ways so that rich multimodal processing of all sorts (including digital video, sound tracks, conversations, and virtual reality sessions) could form the bedrock of a new generation of content management technologies.

However, this potential is far from being realised. Funded projects need to be very carefully chosen for their impact on the local economy and everyday activities.

Threats

One of the biggest threats to Maltese LT is either that it is not done and usurped by English LT, or that certain important use cases and applications are taken over by language-independent development methodologies which do not adequately respect the language-dependent subtleties required for high quality performance. An earlier attempt at the creation of a Maltese spellchecker by a large corporation suffered from this defect and turned out to be completely unusable.

Gaps

First, a large-scale LT R&D programme for Maltese should try to focus on all of the weaknesses mentioned above for the language to increase its DLE score. However, we should bear in mind that the primary target is not merely the DLE score itself, but what the DLE score is supposed to represent: digital support for Maltese, and thus *higher quality* LT, both in its own right, and in relation to other advanced technologies, and in particular AI, alongside which it plays an essential role. The main gaps at present are in three general areas: (i) tools (ii) resources and (iii) support.

[46] See Jocelyne, (2017): http://www.lt-innovate.org/content/language-technology-has-always-been-ai

As far as tools are concerned, the fact is that today we still do not have the barest minimum required for a BLARK (Basic Language Resource Kit), as originally envisaged by Stephen Krauwer in 1998.[47] So we need to focus on not only creating a solid set of building blocks that will serve to build more advanced applications, but also providing ready and universal access to them with the help of platforms like ELG. Minimally, we need to develop the kind of robust parsing machinery and higher level text analytics that are currently regarded as routine for better resourced languages, such as relation and topic detection, sentiment analysis, automated summarisation. The potential for machine translation is beginning to be appreciated thanks to the efforts by the EC but more effort is required locally, beyond the limited timeframe of the NLTP project, for the necessary quality to be achieved in all the domains where it can usefully serve. This requires a concerted policy to facilitate the extraction and refinement of bilingual resources at their point of creation. Voice technology, and particularly ASR (Automatic Speech Recognition), is another priority, since it imparts a highly tangible quality to LT that makes sense to ordinary people in everyday situations. Finally, we lack multimodal tools such as speech-to-speech translation, automated scene description, robot interaction or sign language generation.

These days, almost all of the above tools are driven by machine learning, and thus depend for their quality on the availability of suitable data resources. If we are to have intelligent multimodal, multilingual machinery, we need to have appropriate multimodal and multilingual corpora to train it. This requires a significant effort which, if the resources are to faithfully reflect their inherent regional characteristics, must be developed at national level.

This beings us to the third gap: that of support. Language is so fundamental an element in our society and culture that it pervades all sectors. So it is, potentially, with LT. Yet LT has not received the kind of recognition that is normally afforded to language by various national institutions. If LT for Maltese is to thrive, it needs to be recognised as a national area of priority that requires nurturing, management and support. Most of the language resources and tools that exist today have been created opportunistically. This is a short-term expedient that creates gaps and discontinuities. Language resources and tools need to be commissioned to fit carefully identified needs, and curated on a permanent basis. This requires commitment at national level, and a serious budget, as seen in Spain, Estonia, The Netherlands and even small countries like Iceland Nikulásdóttir et al. (2020). Currently, the institutions that are responsible for the Maltese language adopt a helpful stance towards LT, but have not really taken on board the commitment that is required to ensure that it flourishes and exploits its full potential.

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.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-

[47] http://www.elsnet.org/dox/blark.html

multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520.

Jo-Ann Bajada. Investigation of Translations Equivalences from Parallel Texts. Technical report, Dept CSAI, University of Malta, 2004.

Laurent Bié, Aleix Cerdà-i Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis. Neural translation for the European Union (NTEU) project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 477–478, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.60.

Albert J. Borg and Marie Azzopardi-Alexander. *Maltese*. Routledge, London and New York, 1997.

Mark Borg, Keith Bugeja, Colin Vella, Gordon Mangion, and Carmel Gafa. Preparation of a Free-Running Text Corpus for Maltese Concatenative Speech Synthesis. *Perspectives on Maltese Linguistics, Studia Typologica 14*, pages 297–318, 2014.

Thorsten Brants. TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 224–231, Seattle, Washington, 2000. ACL. URL http://acl.ldc.upenn.edu/A/A00/A00-1031.pdf.

John Camilleri. A Computational Grammar and Lexicon for Maltese. In *MSc Thesis*. Chalmers University, Gothenborg, 2013.

John Camilleri and Michael Rosner. LEXIE: an experiment in lexical information extraction. In *Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains*, pages 19–26, Borovetz, Bulgaria, 01 2009. ISBN 978-954-452-009-0.

L. Cauchi. L'ilsien malti il bieraħ u l-lum, 1994.

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

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. Survey: Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892, December 2017. doi: 10.1162/COLI_a_00302. URL https://aclanthology.org/J17-4005.

Keith Cortis and Brian Davis. A dataset of multidimensional and multilingual social opinions for malta's annual government budget. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):971–981, May 2021. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18120.

Karen Ebert. Aspect in Maltese. In Östen Dahl, editor, *Tense and Aspect in the Languages of Europe*, pages 753–788. Mouton de Gruyter, Berlin, 2000.

Ray Fabri. The Tense and Aspect System of Maltese. In Rolf Thieroff, editor, *Tempussysteme in europäischen Sprachen II*, pages 327–343. Niemeyer, Tübingen, 1995.

Ray Fabri. Maltese. In C. Delcourt and P. van Sterkenburg, editors, *The Languages of the 25. Revue belge de Philologie et d'Histoire: RBPH*, pages 17–28. John Benjamins, Amsterdam-Philadelphia, 2011.

Robert Farrugia. SAMILS – A Semi-Automatic Machine Indexing for Legal Systems. Technical report, Dept CSAI, University of Malta, 2000.

Albert Gatt, Marc Tanti, Adrian Muscat, Patrizia Paggio, Reuben A Farrugia, Claudia Borg, Kenneth P Camilleri, Michael Rosner, and Lonneke van der Plas. Face2Text: Collecting an Annotated Image Description Corpus for the Generation of Rich Face Descriptions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1525.

Albert Gatt and Slavomír Čéplö. Digital corpora and other electronic resources for Maltese. In A. Hardie and R. Love, editors, *Proceedings of Corpus Linguistics*. University of Lancaster, UCREL, 2013.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. MASRI-HEADSET: A Maltese Corpus for Speech Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.784.

Ignasi Badia i Capdevila. A view of the linguistic situation in Malta, 2004. URL https://www.gencat.cat/llengua/noves/noves/hm04primavera-estiu/docs/a_badia.pdf.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1293.

Gordon Mangion. Spelling Correction for Maltese. Technical report, Dept. Computer Science and Artificial Intelligence, University of Malta, Msida MSD2080, 1999.

Paul Micallef. *A Text to Speech Synthesis System for Maltese*. PhD thesis, University of Surrey, 1997.

Manwel Mifsud. *Loan verbs in Maltese: a descriptive and comparative study.* Brill, Leiden etc, 1995.

Ruth Mizzi. The Development of a Statistical Spell Checker for Maltese. Technical report, Dept. Computer Science and Artificial Intelligence, University of Malta, 2000.

Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.418.

Patrizia Paggio, Luke Galea, and Alexandra Vella. Prosodic and gestural marking of complement fronting in Maltese, February 2018. URL https://doi.org/10.5281/zenodo.1181805.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL https://aclanthology.org/P17-1178.

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.

Michael Rosner. Electronic Language Resources for Maltese. In *Proc Bremen Workshop on Maltese Linguistics*. Springer, 2008.

Michael Rosner and Jan Joachimsen. *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012a. ISBN 978-3-642-30680-8. Available online at http://www.meta-net.eu/whitepapers.

Michael Rosner, Joe Caruana, and Ray Fabri. Maltilex: A computational lexicon for Maltese. In M. Rosner, editor, *Computational Approaches to Semitic Languages: Proceedings of the Workshop held at COLING-ACL98, Université de Montréal, Canada*, page 97–105, 1998.

Mike Rosner and Jan Joachimsen. *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 9 2012b. URL http://www.meta-net.eu/whitepapers/volumes/maltese. Georg Rehm and Hans Uszkoreit (series editors).

Roberts Rozis and Raivis Skadiņš. Tilde MODEL - Multilingual Open Data for EU Languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden, May 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-0235.

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.

Cristina Vertan, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Alex Killing, Diane Evans, and Paola Monachesi. Crosslingual Retrieval in an eLearning Environment. In *Proc. Congress of the Italian Association for Artificial Intelligence*, pages 839–847, 09 2007. ISBN 978-3-540-74781-9. doi: 10.1007/978-3-540-74782-6_76.

Andrei Zammit, Claudia Borg, Lonneke van der Plas, and Slavomír Čéplö. A Dependency Parser for Maltese: Comparing the impact of transfer learning from Romance and Semitic Languages, 2019. URL https://www.academia.edu/40156522/A_Dependency_Parser_for_Maltese_Comparing_the_impact_of_transfer_learning_from_Romance_and_Semitic_Languages.

Slavomír Čéplö. Constituent order in Maltese: A quantitative analysis. In *PhD Thesis*. Charles University, Prague, 2018.