EUROPEAN
LANGUAGE
EQUALITY

D1.17

Report on the Greek Language

Authors	Maria Gavriilidou, Maria Giagkou, Dora Loizidou, Stelios Piperidis
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	D1.17
Deliverable title	Report on the Greek Language
Type	Report
Number of pages	30
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	Maria Gavriilidou, Maria Giagkou, Dora Loizidou, Stelios Piperidis
Reviewers	Natalia Resende, Bessie Dendrinos
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
AI4EU	AI4EU (EU project, 2019-2021)
Athena RC	Athena Research and Innovation Center in Information, Communication and Knowledge Technologies
BBT	BackBone Thesaurus
BERT	Bidirectional Encoder Representations from Transformers
CAT	Computer-Aided Translation
CLARIN	Common Language Resources and Technology Infrastructure
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DL	Deep Learning
DH	Digital Humanities
DTB	Digital Transformation Bible
EC	European Commission
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELETO	Hellenic Society for Terminology
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
EMEA	European Medicines Agency
EP	European Parliament
EU	European Union
GDPR	General Data Protection Regulation
HNC	Hellenic National Corpus
HPC	High-Performance Computing
IATE	Interactive Terminology for Europe
ILSP	Institute for Language and Speech Processing
IPR	Intellectual Property Rights
LaBSE	Language-agnostic BERT Sentence Encoder
LR	Language Resource/Resources
LT	Language Technology/Technologies
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing

NLU	Natural Language Understanding
SME	Small and Medium-sized Enterprise
TM	Translation Memory

Abstract

This report is part of the European Language Equality (ELE) reports series that seeks to not only delineate the current state of affairs for each of the European languages covered, but to additionally – and most importantly – identify the gaps and factors that hinder further development in Language Technology (LT). Identifying such weaknesses lays the groundwork for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030. The report at hand sketches the state of affairs for the Greek language, the official language of two EU member states, Greece and Cyprus.

Following a brief introduction to the history, prominent linguistic features, writing system and dialects of Greek, the report focuses on the presence of Greek in the digital sphere; this section discusses the progressive digitisation of the Greek and Cypriot societies, the slow but steady prevalence of Greek-based digital tools and applications that replace previously dominant English-based ones, and the availability of Greek digital public open data. It gives an overview of the status of language resources, and of tools, services and applications concerning Greek; it presents and discusses the variety of language resources for Greek, either intended for human users or supporting LT systems: corpora (general and domain-specific, synchronic and diachronic, monolingual, bilingual and multilingual, parallel, written, spoken and multimodal corpora), lexical/conceptual resources (e. g. computational lexica and online dictionaries, terminological lists and glossaries, thesauri), language models and LT tools and services for the processing of Greek, either for written text or for speech, provided by the R&D community or by the local industry and found at different levels of maturity.

The national policies concerning Artificial Intelligence (AI) and LT are briefly reviewed, in particular with respect to the prevalence of language-centric AI or LT. The Cyprus Strategy for AI and the central role reserved for AI in the Greek Digital Transformation Bible indicate that national policy makers have full understanding of its importance. However, the fact that a vision and specific plans for supporting LT are missing from strategic documents indicates that the fundamental contribution of LT to achieving ubiquitous human-centred AI has not been adequately recognised yet. On the positive side, the establishment of a research infrastructure dedicated to Language Resources coupled with a sister infrastructure dedicated to the Humanities is evaluated as a critical facilitator of LT development.

The presence of AI and LT is also gradually attested to in the academic domain: almost all Greek Universities offer courses (mainly at the postgraduate level) on Natural Language Processing and Language Technology, either as autonomous courses, or coupled with Artificial Intelligence, Big Data and Data Science. With respect to entrepreneurship in the LT and AI domain in Greece and Cyprus, the local industry is small but active; it consists mainly of SMEs offering services both in the country and abroad.

The report concludes that technological support for Greek has progressed in the past decade, while digital language resources have both increased in volume and improved in quality and variety. A critical factor for the availability of resources and tools for Greek has been the creation of Language Resources Infrastructures that cater for storage, curation, and distribution of datasets and technologies/services, properly described with metadata and accompanied by clear and explicit licensing terms.

Despite this progress, when compared to the so-called big languages, Greek is obviously disadvantaged. Prominent among the challenges impeding the development of LT for Greek, is the fact that LT is not included in the overall language policies or in the AI strategies of Greece and Cyprus, while the recognition of the significance of language-centric AI is still lacking. Lack of continuity in research and development funding is an additional progress hampering factor. In conclusion, there is a desperate need for a large, coordinated initiative focused on overcoming the differences in language technology readiness for European languages.

Περίληψη

Η παρούσα έκθεση αποτελεί μέρος της σειράς εκθέσεων του έργου Ευρωπαϊκή Γλωσσική Ισότητα (European Language Equality, ELE), οι οποίες συνοψίζουν τα αποτελέσματα μιας λεπτομερούς εμπειρικής διερεύνησης του επιπέδου τεχνολογικής υποστήριξης των ευρωπαϊκών γλωσσών. Μέσω της σειράς αυτής επιδιώκεται όχι μόνο να περιγραφεί η τρέχουσα κατάσταση για κάθε μία από τις υπό εξέταση γλώσσες, αλλά επιπλέον – και κυρίως – να εντοπιστούν τα κενά και οι παράγοντες που εμποδίζουν την περαιτέρω ανάπτυξη της Γλωσσικής Τεχνολογίας (ΓΤ). Ο εντοπισμός αυτών των αδυναμιών θέτει τις βάσεις για μια ολοκληρωμένη και τεκμηριωμένη πρόταση για τα μέτρα που απαιτείται να ληφθούν, ώστε να επιτευχθεί Ψηφιακή Γλωσσική Ισότητα (Digital Language Equality) στην Ευρώπη έως το 2030.

Η έκθεση σκιαγραφεί την κατάσταση ως προς την τεχνολογική υποστήριξη της ελληνικής γλώσσας, επίσημης γλώσσας δύο κρατών μελών της ΕΕ, της Ελλάδας και της Κύπρου.

Μετά από μια σύντομη εισαγωγή με γενικά στοιχεία για τη γλώσσα (ιστορία, κύρια γλωσσικά χαρακτηριστικά, σύστημα γραφής και διάλεκτοι), η έκθεση επικεντρώνεται στην παρουσία της ελληνικής γλώσσας στην ψηφιακή σφαίρα. Η ενότητα αυτή εξετάζει τη σταδιακή ψηφιοποίηση της ελληνικής και της κυπριακής κοινωνίας, την αργή αλλά σταθερά αυξανόμενη εμφάνιση εργαλείων και εφαρμογών για τα Ελληνικά, που αντικαθιστούν τα προηγουμένως κυρίαρχα εργαλεία για την Αγγλική. Διερευνά τη διαθεσιμότητα ελληνικών ψηφιακών ανοικτών δεδομένων του δημόσιου τομέα καθώς και εκείνων που παρέχονται από την ακαδημαϊκή και ερευνητική κοινότητα.

Στη συνέχεια γίνεται μια επισκόπηση της διαθεσιμότητας γλωσσικών πόρων, εργαλείων, υπηρεσιών και εφαρμογών για τα Ελληνικά: σώματα κειμένων (γενικής γλώσσας ή για συγκεκριμένο θεματικό πεδίο, συγχρονικά και διαχρονικά, μονόγλωσσα, δίγλωσσα και πολύγλωσσα, παράλληλα, γραπτά, προφορικά και πολυτροπικά), λεξιλογικοί / εννοιολογικοί πόροι (π.χ. υπολογιστικά και διαδικτυακά διαθέσιμα λεξικά, ορολογικοί κατάλογοι και γλωσσάρια, θησαυροί), γλωσσικά μοντέλα και εργαλεία και υπηρεσίες για την υπολογιστική επεξεργασία γραπτού κειμένου ή ομιλίας, που παρέχονται από την Ε&Α κοινότητα ή από ιδιωτικές εταιρείες και βρίσκονται σε διαφορετικά επίπεδα ωριμότητας.

Οι εθνικές πολιτικές για την Τεχνητή Νοημοσύνη (ΤΝ) εξετάζονται εν συντομία, ιδίως ως προς το αν εξειδικεύονται σε αυτές στόχοι και μέτρα για τη ΓΤ. Η Εθνική Στρατηγική για την ΤΝ της Κύπρου και ο κεντρικός ρόλος που επιφυλάσσεται για την ΤΝ στην Βίβλο Ψηφιακού Μετασχηματισμού της Ελλάδας δείχνουν ότι οι φορείς χάραξης πολιτικής σε εθνικό επίπεδο έχουν πλήρη κατανόηση της σημασίας της ΤΝ. Ωστόσο, το γεγονός ότι από τα κείμενα εθνικής στρατηγικής απουσιάζει το όραμα και συγκεκριμένα σχέδια ειδικά για την υποστήριξη της Γλωσσικής Τεχνολογίας φανερώνει ότι δεν έχει ακόμη αναγνωριστεί επαρκώς το ότι η Γλωσσική Τεχνολογία είναι κρίσιμο συστατικό για την ανάπτυξη ανθρωποκεντρικής Τεχνητής Νοημοσύνης. Ως θετική εξέλιξη που θα συνδράμει μελλοντικά στην ανάπτυξη ΓΤ για τα Ελληνικά αξιολογείται η δημιουργία της εθνικής ερευνητικής υποδομής Γλωσσικών Πόρων.

Στον ακαδημαϊκό χώρο, η ΓΤ σταδιακά εντάσσεται σε όλο και περισσότερα προγράμματα σπουδών, κυρίως μεταπτυχιακού επιπέδου: σχεδόν όλα τα ελληνικά πανεπιστήμια προσφέρουν μαθήματα για την Επεξεργασία Φυσικής Γλώσσας και τη Γλωσσική Τεχνολογία, είτε ως αυτόνομα μαθήματα, είτε σε συνδυασμό με την Τεχνητή Νοημοσύνη, τα Μεγάλα Δεδομένα και την Επιστήμη Δεδομένων. Όσον αφορά στην επιχειρηματικότητα στον τομέα της ΓΤ και της ΤΝ στην Ελλάδα και την Κύπρο, ο κλάδος είναι μικρός αλλά ενεργός. Αποτελείται κυρίως από ΜΜΕ που προσφέρουν υπηρεσίες στις δύο χώρες και στο εξωτερικό.

Συμπερασματικά, τα τελευταία δέκα χρόνια έχει επιτευχθεί σημαντική πρόοδος ως προς την τεχνολογική υποστήριξη της ελληνικής γλώσσας. Οι διαθέσιμοι ψηφιακοί γλωσσικοί πόροι έχουν αυξηθεί σε όγκο και έχουν βελτιωθεί σε ποιότητα και ποικιλία. Η δημιουργία Υποδομών Γλωσσικών Πόρων, οι οποίοι έχουν τη δυνατότητα αποθήκευσης, επιμέλειας και διανομής συνόλων δεδομένων και τεχνολογιών/υπηρεσιών που περιγράφονται με τα κατάλληλα

μεταδεδομένα και συνοδεύονται από σαφείς και ρητούς όρους χρήσης, έχει παίξει σημαντικότατο ρόλο στη διαθεσιμότητα πόρων και εργαλείων για την ελληνική γλώσσα.

Παρά την πρόοδο αυτή, όταν η ελληνική γλώσσα συγκρίνεται με τις λεγόμενες μεγάλες γλώσσες, βρίσκεται προφανώς σε μειονεκτική θέση. Το γεγονός ότι η ΓΤ δεν περιλαμβάνεται στη συνολική γλωσσική πολιτική ή στον στρατηγικό σχεδιασμό της Ελλάδας και της Κύπρου για την ΤΝ είναι ασφαλώς ένας από τους παράγοντες που δυσχεραίνουν τη μελλοντική ανάπτυξη του πεδίου. Η αναγνώριση της σημασίας της ΤΝ με βάση τη γλώσσα εξακολουθεί να είναι περιορισμένη. Η αποσπασματική χρηματοδότηση έργων Ε&Α που εστιάζουν στην ελληνική γλώσσα και στην υπολογιστική επεξεργασία της δεν διασφαλίζει επαρκώς τους πόρους που απαιτούνται, ώστε το επίπεδο τεχνολογικής υποστήριξης της γλώσσας μας να είναι συγκρίσιμο με αυτό άλλων ευρωπαϊκών γλωσσών. Ελλείψει τεχνολογικής υποστήριξης, οι λιγότερο ομιλούμενες γλώσσες, μεταξύ των οποίων και η ελληνική, θα βαίνουν συρρικνούμενες στην ψηφιακή σφαίρα και θα εκτοπίζονται όλο και περισσότερο από τις κυρίαρχες διεθνείς γλώσσες, έως την ψηφιακή τους εξαφάνιση. Η ανάγκη για μια μεγάλης κλίμακας, συντονισμένη πρωτοβουλία που να επικεντρώνεται στην ενίσχυση της τεχνολογικής ετοιμότητας της ελληνικής γλώσσας είναι επιτακτική.

Ένα τέτοιο πρόγραμμα θα πρέπει πρωτίστως να υποστηρίξει: i) τη συντήρηση, επέκταση και βιωσιμότητα των υποδομών που σχετίζονται με τη Γλωσσική Τεχνολογία, ii) εθνικές ή/και ευρωπαϊκές συντονισμένες δράσεις για την εξασφάλιση της πρόσβασης σε ανοικτές υπολογιστικές υποδομές υψηλών επιδόσεων, iii) συντονισμένες δράσεις για την ανάπτυξη γλωσσικών πόρων μεγάλης κλίμακας, έτοιμων να τροφοδοτήσουν μεγάλα γλωσσικά μοντέλα που υποστηρίζουν ένα ευρύ φάσμα εφαρμογών, iv) στοχευμένες δράσεις για την κάλυψη των παρατηρούμενων κενών σε δεδομένα ομιλίας και πολυτροπικά δεδομένα, v) μέτρα που διασφαλίζουν ότι η σημασία της Γλωσσικής Τεχνολογίας και της ΤΝ αναγνωρίζεται επαρκώς και περιλαμβάνεται στις εθνικές πολιτικές για τη γλωσσική, πολιτιστική και τεχνολογική ανάπτυξη, vi) συντονισμένες δράσεις για την περαιτέρω ενίσχυση του ψηφιακού αλφαβητισμού της ερευνητικής κοινότητας και της κοινωνίας στο σύνολό της, vii) συντονισμένες δράσεις για την προώθηση της κουλτούρας κοινής χρήσης δεδομένων, συμπεριλαμβανομένου του λογισμικού ανοικτού κώδικα, με τη συμμετοχή όλων των ενδιαφερομένων, του δημόσιου τομέα, της έρευνας και της βιομηχανίας.

1 Introduction

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

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

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

[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

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

2 The Greek Language in the Digital Age

2.1 General Facts

Greek is the national and official language of Greece, one of the two official languages of Cyprus and, since 1981, one of the official languages of the European Union. It is spoken as a mother tongue by approximately 95% of the 10.7 million inhabitants of Greece and by approximately 840,000 Greek Cypriots. It is also used by a total of approximately 5 million people of Greek origin, members of Greek communities (the *Diaspora*) worldwide, mainly in the USA, Australia (Melbourne has been named 'the third largest Greek city in the world'), Canada, Europe (UK and Germany mainly), the former Soviet Union countries, Turkey, and Egypt. It is the language used in state institutions, including both lower and higher education, in Greece and Cyprus.

Greek is an Indo-European language, the only surviving member of the Hellenic branch of the Indo-European language family. Unlike Latin, which gave rise to several daughter languages, the only descendant of Ancient Greek is Modern Greek.

The Greek writing system has been the Greek alphabet for most of its history. The Modern Greek alphabet consists of 24 letters. The official orthography of Modern Greek is the simplified *monotonic* (single stress) system, which utilises only stress mark and diaeresis. The traditional system, called the *polytonic* (multiple stress) system, is still used internationally for the writing of Ancient Greek.

Greek is a heavily inflectional language, and has an extensive set of derivational affixes, whereas the system of compounding is relatively limited, but productive. As regards syntax, Greek presents a free word order, the neutral word order being Verb-Subject-Object or Subject-Verb-Object. This allows the speakers to form utterances in a wide variety of ways putting the focus on different parts of the sentence. The rich case system makes free word order possible and offers crucial information to syntactic analysis: nominative case is used only for subjects and predicates, and accusative for objects of most verbs and of many prepositions, genitive for possessives and for objects of some verbs and prepositions. Consequently, recognition of syntactic roles is more straightforward than in languages with no cases. Two significant features of the Greek vocabulary are extent and word length. One reason for the size of the vocabulary is the great number of synonyms observed. The abundance of synonyms is due to their origin from the various dialects, besides loanwords from other languages. Another reason for the extensive vocabulary is the productivity of the derivational morphological system. As regards word length, Greek has very few one-syllable words. Two- or three-syllable words are the majority, but multi-syllable words are not rare at all (even eight or nine-syllable words).

Dialects and minority languages

Almost all Modern Greek varieties descend from the *Koiné* (Browning, 1969), the common supra-regional form of Greek spoken and written during the Hellenistic period, the Roman Empire and the early Byzantine Empire. After World War II, the various Greek dialects gradually declined and some (e. g. the Cappadocian dialect, the Tsakonian dialect or Grico, the Greek dialect spoken in a handful of villages in southern Italy, an area also known as Magna Grecia) are considered practically extinct. The dialects and languages used by minority populations are considered as elements of cultural identity and tradition. The modern way of living, urbanism, the use of the standard variety in education, administration and the mass media has led progressively to their exclusive use in their respective communities. The most identifiable dialects of Greek which are in use to date are the Pontic dialect (variety of Modern Greek originally spoken in the Pontus area, and today mainly in northern Greece), the Cretan dialect (spoken mainly in Crete) and Cypriot Greek (spoken mainly in Cyprus).

The only language which is recognised as a minority language is Turkish, which has the status of minority language in Thrace (a region in Northeastern Greece). However, there are several other languages spoken by minority populations such as the Pomaks (a muslim minority who speak Pomak, a Bulgarian dialect), the Roma (who speak Greek Romani), Slavomacedonians (who speak a Greek variety of Slavomacedonian, which is the national language of North Macedonia), Arvanites (who speak Arvanitika, a dialect of Albanian) and Vlacks (who speak Vlack also known as Aromanian). There are several other languages spoken by ethnic communities in Greece – communities of first, second or third generation immigrants. Some of these are even taught as heritage languages in complimentary schools, including Albanian, Armenian, Arabic, Bulgarian, Russian, Georgian, Ukranian, Serbian, Farsi, Chinese. As for Cyprus, which is still today a member of the Commonwealth and where the official languages are Greek and Turkish – the vernaculars being Cypriot Greek and Cypriot Turkish, there are two recognised minority languages: Armenian and Cypriot Arabic. For the latter, there is a serious effort, funded through the Cypriot Ministry of Education, for its revitalisation. Moreover, given the country's colonial past, since Cyprus was under British rule until the mid-20th century, English is very widely spoken. Other languages spoken by minority populations include Russian, Bulgarian, Romanian, Ukrainian, Croatian, Albanian, Macedonian, Montenegrin, Slovene, Serbian, Bosnian, Polish, German and Hungarian.

2.2 Greek in the Digital Sphere

The Greek and the Cypriot are digitised societies, with Cyprus being more advanced in terms of internet penetration in everyday life. According to data from Eurostat,[3] 77% of individuals in Greece and 91% in Cyprus used the internet in 2021 at least once a week. The percentages were raised to 98% and 99% (Greece and Cyprus respectively) but only for young people aged 16-24. The percentages are lower for individuals aged 25-54, i.e. 91% and 98% respectively, and even lower for individuals aged 65-74, i. e. 32% and 58% respectively.

A survey by the Greek Statistical Authorities, spanning across the first quarter of 2019[4] reveals that 8 out of 10 households have full access to broadband internet at home, while approximately 83% use their smartphones to access the Web. The percentage is higher in Cyprus as, according to the Statistical Service of Cyprus, 93% of the households had Internet access in the first quarter of 2021.

The Greek domain (.gr URLs) has more than 500,000 registered addresses,[5] while the Cyprus domain (.cy URLs) currently has more than 20,000 registered active addresses.[6] The main language of websites under the .gr and .cy domains is Greek, while many of them are bilingual, providing both Greek and English versions. The overall expansion of the Greek digital world and its users have led to intense localisation of major applications for computers and smart phones, and the development of native Greek applications and free software. Greek is now used in all social media, e-government and all Greek companies' online shopping platforms. In international or multinational companies or platforms, Greek is still served mainly through Machine Translation (MT).

[3] https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_ci_ifp_fu&lang=en, last update: December 16, 2021

[4] Survey of the Use of Information and Communication Technologies by Households and People: Year 2019, 8/11/2019 (Available only in Greek: - Έρευνα Χρήσης Τεχνολογιών Πληροφόρησης και Επικοινωνίας από Νοικοκυριά και Άτομα: Έτος 2019 through http://www.statistics.gr)

[5] On April 2021. Source: https://www.eett.gr/opencms/opencms/admin/News_new/news_1479.html

[6] https://www.nic.cy

Sources for Greek open language data on the Internet

Despite the extensive use of Greek in digital communication between citiens, private businesses and public administration, the amount of digital open government language data has not increased significantly. Although all Public Administration's acts and decisions and Open Government Data are shared through dedicated Open Government portals in Greece,[7] language data is either shared in non-processable formats (e. g. in PDF) or not at all. The value of public language data generated and held by the administration has not yet been recognised and no dedicated efforts have been implemented for its management and sharing. In what concerns encyclopedic, educational and other related digital data, the valuable source of Wikipedia is of moderate assistance to Language Technology as regards Greek. While the English Wikipedia contains more than 6 million articles, only approximately 200,000 articles are included in the Greek version.[8] The Greek Wikipedia initiative started in December 2002, but it appears that the contributors' team has not grown considerably, nor is content frequently added (last change recorded on 31.12.2019).[9] The Greek Open Technologies Alliance[10] is actively promoting the expansion of Greek Wikipedia with a range of activities.

Fotodentro,[11] the National Educational Content Aggregator for Primary and Secondary Education, is the central digital service for the distribution of digital educational content to schools. Fotodentro promotes the use of open educational resources, implementing the national strategy for digital educational content.

The *mycontent* initiative[12] of the Open Technologies Organization[13] aims to contribute to free access to information and freely available content. It provides links to open data collections, prominent among which is the Open Library;[14] it currently includes 11,000+ free Greek e-books, 1,500+ audio-books and is enriched daily with new ones.

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 because understanding language is a very complex task; it requires understanding the relationship between words, used in different types of texts (genres) and in different situational contexts, as well as to what the words refer to. To understand these relationships, one needs to have textual, contextual and what is often called "world knowledge". Depending on text and context, messages containing similar information can be lexicalised in different ways and create different socially purposeful meanings.

The computational processing of human languages has been established as a specialised field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, *Language Technology* (LT). While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably Artificial Intelligence (AI)), mathematics and psychology among others.

[7] Transparency portal (https://diavgeia.gov.gr) and Greek Open Data portal (https://www.data.gov.gr)

[8] Source: https://meta.wikimedia.org/wiki/List_of_Wikipedias (accessed 11/01/2022).

[9] Source: https://el.wikipedia.org

[10] https://gfoss.eu

[11] http://photodentro.edu.gr

[12] https://mycontent.ellak.gr

[13] https://eellak.ellak.gr

[14] www.openbook.gr

[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).

In practice, these communities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

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

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in Machine Learning (ML), 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 revolutionising the way in which LT tasks are approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of Artificial Intelligence (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, 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.

- **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 realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just a few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance, for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[16]

4 Language Technology for Greek

This section provides a brief overview of the status of the language resources and tools/services and applications concerning the Greek language.

4.1 Language Data

Monolingual corpora

Several general domain monolingual text corpora serve as material that is representative of contemporary language use: most notably, the Hellenic National Corpus[17] (Gavrilidou, 2002), which contains texts dated from 1990 onwards, with a total of approximately 100 million words, the corpora of the Centre for the Greek Language,[18] which have 7 million words sourced from newspapers, and the Corpus of Greek texts of the University of Athens[19] (Goutsos, 2010) containing 30 million words of texts from 1990 to 2010. This corpus will be enriched by the addition of the Greek Corpus 20[20] (Goutsos et al., 2017), a diachronic corpus of Greek in the 20th century of a total aimed size of 20 million words.

As regards spoken Greek, two collections of authentic everyday conversations between students collected by the University of Athens are available through CLARIN:EL: (a) the first one includes 618 individual archives with a total of more than 1.7 million words of written transcripts of conversations between 2001 and 2006 (the original sound archives are not available) and (b) a corpus including students' conversations in 2020, including both audio

[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://tinyurl.com/2p9ed6tp). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

[17] http://hnc.ilsp.gr

[18] https://www.greek-language.gr/greekLang/modern_greek/tools/corpora/index.html

[19] http://sek.edu.gr

[20] http://greekcorpus20.phil.uoa.gr

recordings and written transcripts. Learner corpora are provided by the Aristotle University of Thessaloniki.

However, all these resources are primarily intended for human users, and are accessible through specially designed interfaces providing concordances, sentence-based data, and statistics, to be used mainly for linguistic research, or for lexicographic or teaching purposes. Large datasets, freely accessible to all (research or industry) for building language models or for developing and testing language processing tools and applications are scarce. Nevertheless, several collections do exist, reflecting general language or specific domains (medical, legal, economic, humanities, etc.), text types (journalistic articles, literature, press releases, political party manifestos, subtitles, etc.), from a variety of sources from the web and adhering to specific selection criteria (e. g. Parliamentary data, COVID-19 related data, Twitter data focusing on verbal aggression and xenophobia, newspaper data for event extraction, blogs, sports magazines, etc.), which have been collected by both manual and automatic methods, in particular focused web crawling. Clearly, Greek has progressed in terms of coverage of specific domains as well as of the general language, but the size of available corpora still does not suffice for valid synchronic linguistic research and cannot guarantee the development of robust and well-performing language processing tools that address contemporary needs.

Bi- and multilingual text corpora

Multiple bi- and multilingual text corpora include Greek as one of the languages, as attested to in three language resources infrastructures, ELRC-SHARE,[21] ELG[22] and CLARIN:EL.[23] Parallel corpora such as those sourced from EU institutions (European Commission, European Parliament, several EU Agencies, e. g. EMEA) have been extensively used for the development and training of MT systems, or to power the translation memory databases of Computer-Aided Translation (CAT) tools. They are mostly available in two versions: either as multilingual corpora (with all language variants) or as bilingual language pairs, raw and sentence-aligned. Besides this, Greek is featured in some multilingual collections covering also non European languages such as Nepali and Hindi. Few domain-specific collections of parallel and/or comparable corpora exist, and that ones that do, mainly cover domains such as Law, Banking, Medicine (especially COVID-19), etc. Multilingual corpora are developed mostly automatically by leveraging automatic web crawling techniques, an approach that has been extensively used by the European Language Resource Coordination[24] initiative (Papavassiliou et al., 2018). Greek is included as one of the languages in several other parallel corpora created by web crawling; most of them are available as parts of the OPUS sub-corpora (Tiedemann, 2012), Wikimatrix, OSCAR and Paracrawl to name but a few.

Multimodal corpora

Some multimodal resources, i. e. corpora and lexica including two or more modalities, e. g. audio combined with text, video with text, and image with text, are available for Greek. Very significant multimodal resources in this category include the multilingual sign language corpora and lexica (Goulas et al., 2018; Efthimiou et al., 2012; Dimou et al., 2012; Adaloglou et al., 2020), an important resource type for the creation of multimodal applications such as Sign Language Recognition and Generation, but also for continuous speech recognition systems for Greek, applications for those who are hard of hearing, etc. Recent attempts to construct multimodal language resources for speech pathology applications are also noteworthy (Varlokosta et al., 2016; Kasselimis et al., 2020).

[21] https://elrc-share.eu

[22] https://live.european-language-grid.eu

[23] https://inventory.clarin.gr

[24] https://www.lr-coordination.eu

Lexical/conceptual resources

As regards lexical/conceptual resources, there exists one large monolingual computational lexicon (LEXIS)[25] containing around 70,000 entries at the morphological level, 45,000 entries at the syntactic level and 30,000 entries at the semantic level, developed manually by lexicographers, which, however, does not seem to have been extensively used yet. With regards to bi-/multilingual reference lexical resources, Greek participates in various resources: the COLLINS multilingual database, IATE (Interactive Terminology for Europe) which is the EU's terminology database, the BackBone Thesaurus, a multilingual thesaurus of digital humanities specialising in archival and library science. More advanced multilingual lexical resources including Greek have been created, such as WordNet, sentiment lexicons, ontologies/semantic networks (ConceptNet).

Lexical resources for educational purposes, e. g. learners' dictionaries and wordlists, have also been developed. For instance, the KELLY project resulted in the creation of monolingual and bilingual wordlists covering 36 language pairs in total (Greek included), supporting foreign/second language learning (Kilgarriff et al., 2014). However, basic resources (e. g. the basic vocabulary for various levels of Greek as L1 or as L2) are still lacking. Online Greek monolingual and bilingual dictionaries for human users are made available by the Centre for the Greek Language. The Greek Terminology society (ELETO) and the Department of Foreign Languages, Translation and Interpreting in Ionian University produce a multitude of terminology lists and glossaries. They, together with CLARIN:EL, have produced and made available many mono- or bilingual terminology lists in many special, diverse domains; indicatively, agriculture, management, literature, linguistics, psychology, library science, immigration – to name but a few. Also, valuable was the contribution of the OROSSIMO project, which commissioned the creation of bilingual term lists (Greek – English) for many scientific and technical domains, such as engineering, medicine, law, astronomy, biology, air traffic, computer science, etc. These domain-specific lexical resources are valuable as authority lists for the training and/or evaluation of MT systems.

Models and grammars

Greek features in some multilingual and/or monolingual language models; indicatively, models catering for the task of sentence boundary detection for multiple languages, of deep contextualised word representation (trained for 44 languages), Greek domain specific n-grams (for words and word/tag/lemma tuples of the Environment and Legal domain). Recently, three BERT (Bidirectional Encoder Representations from Transformers) models have been developed for Greek: (a) a syntax-augmented Multilingual BERT trained on four NLP tasks, including text classification, question answering, named entity recognition, and task oriented semantic parsing (Ahmad et al., 2021), (b) a language-agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2020), which is a BERT-based model trained for sentence embedding for 109 languages and (c) GreekBERT, a Greek monolingual version of the BERT pre-trained language model created by the Athens University of Economics and Business (Koutsikakis et al., 2020).

4.2 Language Technologies and Tools

During the last few years, progress has been shown with regards to the coverage of Greek in terms of LTs. Existing basic NLP tools have been improved by adopting deep learning methodologies and algorithms, as well as new technologies such as neural networks. Preprocessing tools catering for conversion between formats, language identification, align-

[25] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-6105-D

ment of parallel texts have been in place for quite some time, as well as NLP support operations for Greek; indicatively, Ellogon, a multilingual, cross-platform, general-purpose text engineering environment (Petasis et al., 2002), the NLP toolkit of the Athens University of Economics and Business[26] and the ILSP pipeline (Prokopidis and Piperidis, 2020). The basic ILSP pipeline includes tools for various types of annotation, i. e. sentence splitting, tokenization, POS tagging, lemmatization, chunking, and dependency parsing. Lately, a new pipeline (or toolkit) has been developed based on neural network technology. All pipelines are available for use through the ELG and the CLARIN:EL infrastructures. Tools for more advanced tasks such as monolingual Information Extraction, Event Detection and Named Entity Recognition have also improved over the last few years, by being trained on new datasets and applied to a variety of domains. Various other applications such as anonymisation, NLG and sentiment analysis can be found at different levels of robustness and completeness. Concerning multilingual text processing, MT systems have significantly improved their coverage of Greek: eTranslation, Google Translate and DeepL successfully treat Greek, both as source and as target language (Greek as target language poses greater difficulties than as source, mainly due to its rich inflection). A number of MT systems have also been developed by smaller companies in Greece and other EU Member States, and by academic and research organisations.

Speech processing (recognition and synthesis) have seen important progress: dictation systems for Greek with a number of domain specific implementations have been made available by commercial providers; high-calibre multi-language, including Greek, speech synthesis technologies have also been developed by an ILSP spin-off company which was acquired by a big multinational company in 2017. This example seems to verify a trend evidenced in Europe, that European start-ups that develop advanced technologies are often acquired or merged with big multinational commercial organisations, which results in what is often described by the LT community as a loss of European assets. Greek speech synthesis, with varying quality in terms of intelligibility, naturalness and expressivity of the generated speech, is also part of the portfolios of some multinational commercial providers. Several Greek-speaking digital assistants are also currently available; Theano,[27] the chatbot/conversational assistant developed by Athena RC with state-of-the-art AI technologies, specialises in COVID-19 and provides relevant information to citizens.

Besides national organisations' R&D endeavors, Greek is progressively included in tools and technologies catering for many languages; thus, Greek is present in speech translation technologies (e. g. iTranslate) and in assistive software for visually impaired people that turns printed text into speech or into Braille (e. g. KNFB Reader, RoboBraille). However, given that Greek is one of the lesser-used and taught languages, the market for Greek products is limited and, thus, international companies are disinterested in investing in them. Therefore, the transition to oral human-computer interaction in Greek is still underdeveloped.

4.3 The case of Cypriot Greek

Cypriot Greek is a dialectal variety of Modern Greek spoken in Cyprus and by Cypriot Greek diaspora. Standard Modern Greek, as the official language of Cyprus, is the medium of education and the language of the Cypriot media. Cypriot Greek is used mainly in oral speech and in specific written speech types such as poetry and literature (Karyolemou, 2001). The Cypriot Greek dialect – the lingua franca in most multicultural Cypriot communities (Arvaniti, 1999) – signifies the island's historic past and the many conquests it has endured over the centuries; the influence of various languages such as Latin, Venetian, Medieval French, Catalan, Arabic, Turkish and English, is remarkable in the Cypriot Greek dialect.

[26] https://github.com/nlpaueb/gr-nlp-toolkit

[27] https://www.athenarc.gr/en/theano-covid19-chatbot

From a sociolinguistic perspective, although Greek Cypriots are proud of their dialect, they consider Modern Greek "superior", "more beautiful" and "more correct" than Cypriot Greek (Sciriha, 1996). Cypriot Greek is often not easily understood by speakers of Standard Modern Greek (Arvaniti, 1999) and so a Greek Cypriot speaker, in the presence of a Standard Modern Greek speaker, switches to a provincial variety of Modern Greek (Karyolemou, 2001). Despite the use of the Standard Modern Greek language by Greek Cypriot speakers, several features of the Cypriot Greek dialect prevail.

Cypriot Greek is distinguished from Standard Modern Greek on different levels of the linguistic system: phonology and phonetics, grammar, vocabulary, morphology, syntax, pragmatics, semantics and orthography (Terkourafi, 2007). Two main features of Cypriot Greek which differ from Standard Modern Greek are the following:

1. The use of gemination, "glide hardening" (Armosti, 2009). The sound system contains geminates and palato-alveolar consonants, not represented in the Greek alphabet. There have been discussions to add diacritics for Cypriot Greek.

2. The use of loanwords of many different languages (e. g. muhtar – mouχtaris – γitonia – neighborhood) (Pavlou, 1994).

Taking into consideration all the above, it is often the case that LTs trained on Standard Modern Greek data, fail to appropriately process Cypriot Greek. This is particularly relevant for speech processing technologies; for instance, the performance of any speech recognition system for Standard Modern Greek will decrease significantly when used by a Cypriot speaker. This is also true for written Modern Greek in Cyprus, especially in the legal and public administration domain. Cypriot legalese documents, despite being written in Modern Greek, present a number of differences with respect to vocabulary and terminology use. Hence, an MT system's performance decreases for Cypriot legalese documents. In order to protect this dialectal variety of Modern Greek, as well as the heritage and culture of its speakers, LT research should specifically treat Cypriot Greek.

Language resources and tools/services specifically for Cypriot Greek are sparse. They are mainly general use lexical resources (dictionaries, glossaries and wordlists). Indicatively, some of the most sizeable are the following: *Wikipriaka*, an online dictionary (downloadable and printable) with 2,353 entries. *CySlang* (Κατσογιάννου and Χριστοδούλου, 2019) is an online multimedia glossary of Cypriot Greek slang. It exclusively includes vernacular, youthful and slang words and expressions of modern spoken Cypriot Greek. The *Lexicological database of the Cypriot dialect* is an online dictionary of spoken Cypriot Greek. It contains approximately 15,000 entries and each one is accompanied by an audio file with a pronunciation. The *Electronic dictionary of Cypriot Greek-Portal for the Greek Language* is a subset of an online dictionary including only dialectal forms of Cypriot Greek. Finally, *HelexKids* is a word frequency database that was developed for Greek/Cypriot children from the first to the sixth grade of primary education. The database is based on a corpus of 1.3 million words extracted from 116 textbooks covering a wide range of readers from a variety of topics, from mathematics to physical education.

With respect to corpora, the Cypriot Greek sub-corpus of the *Multi-CAST* corpus (Hadjidas and Vollmer, 2015) comprises three texts (syntactically annotated traditional narratives) which were originally recorded in the 1960's. A recent noteworthy dataset for Cypriot Greek is the Multilingual Corpus of the DIALLS project (DIalogue and Argumentation for Literacy Learning in Schools). It consists of a set of transcripts of classroom interactions of students from ages 5 to 15 years old in seven countries (UK, Portugal, Germany, Lithuania, Spain, Cyprus and Israel). The Cypriot Greek part consists of 19 transcripts.

Finally, as regards tools, the Greek Dialect Classifier (Sababa and Stassopoulou, 2018) identifies Greek text as Cypriot Greek or Standard Modern Greek. It has been trained on a corpus of Cypriot Greek and Modern Greek Facebook posts and tweets.

4.4 Projects, Initiatives, Stakeholders

National policies for AI/LT

The main recent public policy document for AI in Greece is the Digital Transformation Bible (DTB),[28] which outlines the basic principles, the strategic axes and the horizontal and vertical interventions for the digital transformation of Greece. Key reforms critically related to AI are the digitisation of the public sector, interoperability of IT systems and quality services to businesses and citizens, training, upskilling and reskilling of public employees (with an emphasis on digital skills) and digitisation of the health, education and social inclusion of vulnerable groups. The reform aims to develop a holistic framework that will bring technological advancements (i. e. Cloud computing, Business Intelligence, AI, ML, etc.) into the public administration. NLP (together with Information Extraction, ML, Text Mining and Big Data Analytics) is mentioned in the DTB in relation to the objectives of automatic codification of legislation and facilitation of search through the Transparency Portal. It might be the first time NLP is mentioned in a strategic document as a method/technology supporting public administration.

Many of the objectives set by the DTB have been met, and digital services offered to citizens were digitalised quickly and efficiently; additionally, the coronavirus pandemic in Greece acted as an accelerator for the digital transformation of the country. The pandemic offered a perfect opportunity to the education sector for developing new (or deploying existing) digital educational material, and enhancing the digital literacy of educators, however, unfortunately this opportunity was sadly missed.

The development of the National Strategy of AI was initiated in 2020. The final version of the strategy was due to be published in early 2021 but at the time of writing (January 2022) this document had not yet been circulated. The National AI strategy of Cyprus, published in January 2020, focuses on, among others, improving the quality of public services through the use of digital and AI-related applications; creating national data areas and developing ethical and reliable AI. It is worth noting that NLP and LT are not explicitly mentioned in the Strategy and no specific support action is anticipated. The report, however, clearly enumerates some language technologies as examples of AI applications, thus implicitly recognising the vital role of language-centric AI, even if this is not stated as such.

Staying on the subject of national policies, the promotion of research and innovation (as defined in the Greek Recovery and Resilience Plan[29] foresees actions to promote basic and applied research, by

- upgrading the computing infrastructure of the 14 public Research Centers in Greece;
- funding projects in basic and applied research as well as flagship research projects in challenging interdisciplinary sectors with practical applications in Greek Industry;
- the development of the “ELEVATE Greece” platform for the support and global promotion of national startups. The ultimate objective is to expand these services to the entire national innovation ecosystem (i. e. research centres, innovation clusters; competence Centres and highly innovative companies).

In this framework, a significant initiative was launched recently (end of December 2021): the creation of ARCHIMEDES, a new research unit on Artificial Intelligence, Data Science and algorithms. Founded as an independent unit of Athena RC, ARCHIMEDES is an emblematic initiative for Greece whose main goal is to serve basic and applied research in AI in collaboration with universities both in Greece and abroad. In the field of entrepreneurship, the

[28] https://digitalstrategy.gov.gr/principles_of_implementation

[29] More details can be found in the Greek Recovery and Resilience Plan (https://greece20.gov.gr/en/the-complete-plan)

Unit additionally aims to facilitate the transfer of research results to the market. The fact that the significance of LT has attracted the attention of policy makers is proved by the establishment of a research infrastructure dedicated to LT/LRs coupled with a sister-infrastructure dedicated to the Humanities. This will be detailed in the next subsection.

Dedicated language data sharing infrastructures

There are several language resources repositories and research infrastructures in Greece, stemming from R&D activities and initiatives related to Language Resources and Technology, either national or European. The CLARIN:EL[30] infrastructure (Piperidis et al., 2017), coordinates a distributed network of 14 nodes, and collects, stores and distributes language resources and technologies through its inventory of language resources[31]. It also undertakes training and awareness activities on the significance and use of LT. CLARIN:EL is enlisted in the National Roadmap for Research Infrastructures of Greece[32] and is the Greek part of the CLARIN ERIC European Infrastructure. CLARIN:EL services are offered to all Greek users as well as to all users-members of the CLARIN ERIC European Infrastructure. CLARIN:EL, jointly with its sister-infrastructure DARIAH-DYAS[33] (dedicated to the Arts and Humanities) constitute APOLLONIS,[34] the national infrastructure that supports and promotes digital humanities and arts, and LT and innovation in Greece.

The Institute for Language and Speech Processing (ILSP) of the Athena Research Centre has also played an instrumental role in a number of EU wide infrastructures, starting with META-SHARE (Piperidis, 2012), and continuing with ELRC-SHARE (Piperidis et al., 2018) and more recently with the European Language Grid (Rehm et al., 2020). Language resources and technologies for Greek (but also other languages) are shared by the European Language Grid Platform,[35] which aims at listing datasets and language technology services as well as relevant stakeholders (technology development, research centres, small and medium-sized companies and large enterprises) and projects. Finally, the European Language Resource Coordination (ELRC)[36] manages, maintains and coordinates the collection of MT related language resources in all official languages of the EU and CEF associated countries.

LT industry

With respect to entrepreneurship in the LT and AI domain in Greece and Cyprus, the local industry is small but active, offering services in both countries as well as abroad. It consists mainly of SMEs, some founded by young researchers as spin-off companies of research institutions or even as not-for-profit organisations. Their number is fluctuating, as some of them get acquired by big international companies or, unfortunately, need to cease their operations. Currently (January 2022), we estimate that approximately 15-20 SMEs are active in LT in Greece and Cyprus, providing various LT-related services, indicatively: AI, LT (event detection, basic NLP, lexical resources and terminologies), MT and Localisation, Speech Processing (mainly recognition), Data Science/Big Data Analytics. The services offered range from fully developed online platforms to customer specific solutions and to free cutting-edge information technology services. A great variety of applications based on LT and AI is observed: textual data AI-based analysis for the provision of insights to business-related questions, online spellers and grammar checkers, dictionaries and thesauri, speech-to-text products and

[30] https://www.clarin.gr
[31] https://inventory.clarin.gr
[32] http://www.gsrt.gr/News/Files/New987/road-map-web_version_final.pdf
[33] https://dyas-net.gr/dariah-gr-dyas
[34] https://apollonis-infrastructure.gr
[35] https://www.european-language-grid.eu
[36] https://lr-coordination.eu, https://www.elrc-share.eu

conversational AI speech recognition, localisation platform that aids the translation of website or product content, tools to aid CAT/TM users and intelligent personal assistants.

Research/academia

In the academia/research domain, while in the past research was confined to theoretical language and linguistic studies, more and more cells of LT research and development gradually emerge. The vast majority of Greek Universities now offer courses at the postgraduate level in NLP and LT, either as autonomous courses, or coupled with AI, Big Data and Data Science in general.[37] In addition, in many Greek universities and research institutions, LT-related labs have been formed, either generic or with a particular domain focus e. g. LT in the legal domain or in the Greek public administration.

The main research/academic entities in Greece that conduct R&D activities in LT are: the Institute for Language and Speech Processing of the Athena Research Centre, the Department of Informatics of the Athens University of Economics and Business (AUEB), the Institute of Informatics & Telecommunications of the National Centre for Scientific Research "Demokritos", but also almost all Informatics and Linguistics departments in Greek universities conduct some form of LT R&D, with different degrees of focus and intensity.

Cypriot universities and research labs are gradually starting to conduct research on AI but not yet on LT. Academic labs, such as the KIOS Research and Innovation Centre of Excellence at the University of Cyprus, Cyprus University of Technology's Software Engineering and Intelligent Information systems research lab and University of Nicosia's AI Laboratory, are leading the path in the broader AI field in Cyprus.

With respect to language policy, the Centre for the Greek Language[38] is acting as a cooperating, advisory and planning body of the Greek Ministry of Education on matters of language policy and it is the official certification body of attainment in Modern Greek. Multilingualism, linguistic and cultural diversity are the focus of research labs, such as the Greek Language and Multilingualism Laboratory (University of Thessaly)[39] and the Centre for Excellence for Multilingualism and Language Policy (University of Athens).[40] The latter has launched the first Multilingualism Observatory in Greece to investigate multilingualism and language teaching, learning and assessment in Greece, as well as Greek Studies abroad.

Funded projects

Although in the last ten years there has been no funding programme specifically supporting LT in Greece, advancement of LT for Greek has been achieved mainly through the participation of these organisations in national and European funded projects in the broader field of LT and AI. To name but a few, Greek organisations participate in AI4EU, ManyLaws, ELG: European Language Grid, MyDataStories, MOBOT: Intelligent Active MObility Aid RoBOT integrating Multimodal Communication, HumanE-AI-Net: HumanE AI Network, UNBIASED: Fact-provisioning and bias estimation tools for public inoculation against disinformation campaigns, ConvAI – Context-aware abusive language detection in online conversations and many more.[41]

[37] A non-exhaustive list of postgraduate courses can be found in the Digital Humanities Course Registry

[38] http://www.greeklanguage.gr

[39] http://greeklanglab.pre.uth.gr

[40] https://cem.uoa.gr

[41] This list is only indicative and by no means exhaustive. Full lists of the projects can be searched on the above organisations' websites.

5 Cross-Language Comparison

The LT field[42] as a whole has evidenced remarkable progress during the last few 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).

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[43] 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:

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

[43] 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.

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[44]

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[45] 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 –currently in progress– development of a 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.[46]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

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

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

[44] 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.

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

[46] 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 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g. German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[47] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

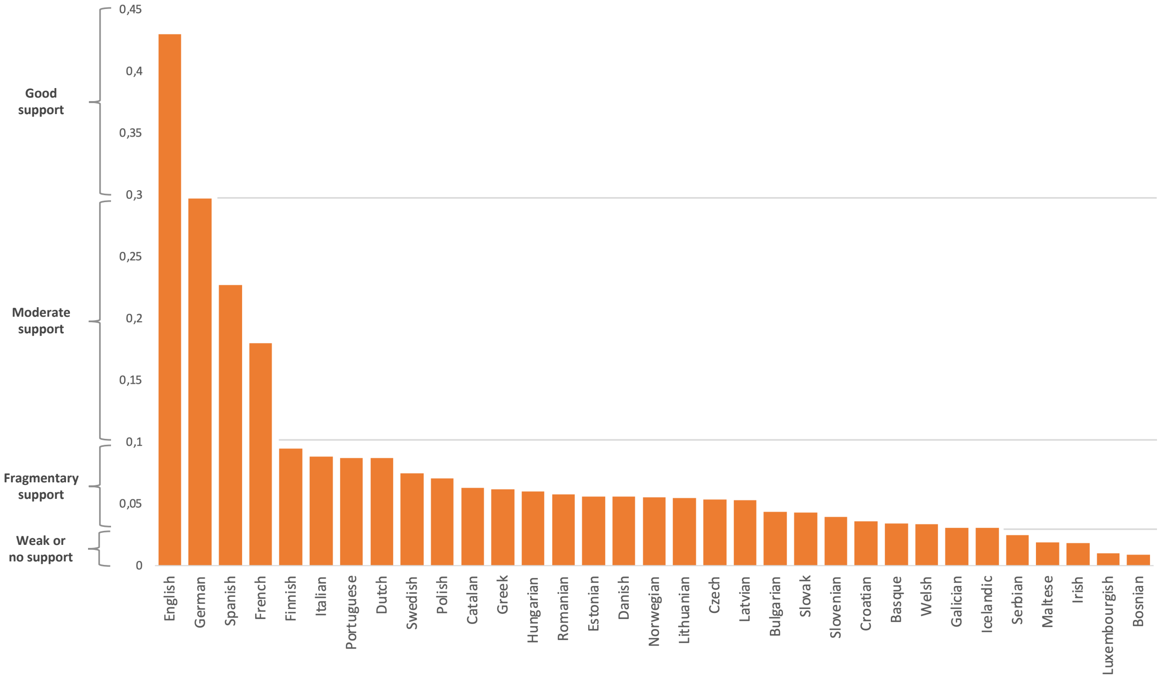

Figure 1: Overall state of technology support for selected European languages (2022)

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,

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

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 the 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e.g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

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

Technological support for Greek has progressed overall in the past decade compared to the state of affairs described in the META-NET White Paper (Maria Gavrilidou et al., 2012). Digital language resources have both increased in volume and improved in quality and variety.

Resources and basic NLP tools are provided by academia, research centres and private companies as outputs of various endeavours (research projects conducted by academic institutions, funded by EU or national funds, commercial projects or self-funded) and made available under various licensing conditions (freely distributed, only for research, only in samples, available through interfaces, etc.). Significant progress has been made with respect to available corpora and lexica, language models, text processing tools, MT and speech processing (synthesis and recognition). The available datasets come from a variety of sources and they cover several thematic domains, text types and languages; raw or annotated, monolingual, bi- and multilingual. However, their size is lagging behind in terms of appropriateness for building really large language models or robust, ready to use tools and applications.

A critical factor that benefited the overall availability of resources and tools for Greek has been the creation of Language Resources Infrastructures that cater for storage, curation, and distribution of datasets and technologies/services, appropriately described with the relevant metadata and accompanied by clear and explicit licensing terms. Furthermore, the LRs infrastructures have actually promoted the openness and sharing culture among researchers and developers. Several organisations have been moving forward as regards the digitisation of their services and workflows and are keen to make their data openly available. However, this trend is still not sufficiently widespread.

Despite the attestable progress, when comparing Greek to the so-called big languages, the abysmal difference in terms of quantity, size and quality of resources and tools is evident. Moreover, while looking at more advanced datasets and tools, Greek is severely disadvantaged. In this respect, efforts in the coming years should be concentrated on the further development of large-scale monolingual corpora that can be used for training massive language models like GPT-3. Semantically annotated datasets, semantic lexica and knowledge bases, and datasets that can be used for anonymisation, simplification, summarisation, text levelling and question-answering systems should also be prioritised. Speech and multimodal data are scarcely available, if at all, limiting the potential for the development of conversational agents, among others. Greek is dramatically deprived particularly when it comes to conversational data or speech in informal settings that is generated by speakers of different ages, genders and linguistic/dialectal backgrounds. The transition to ubiquitous human-computer interaction in Greek, supported by state-of-the-art research results in Natural Language Understanding and Generation is, unfortunately, still far away.

Further challenges posing impediments to the development of LT for Greek include:

- Scarcity of data: as Greece and Cyprus are small countries, the production of digital (language) data is limited, especially when compared to larger countries with broadly used languages at the national and international levels;
- Lack of experience in the use of LT: the deployment of digital tools and methods in many disciplines, including life sciences and humanities, has only recently been introduced. Researchers and professionals in domains other than LT are still to be convinced about its benefits;
- Issues related to IPR or GDPR render resource owners hesitant about sharing their datasets. Non-explicit, unclear distribution and use terms restrict sharing, use and repurposing of digital texts and language processing tools. The majority of resources – when made available – pose restrictions on the types of uses they allow (for example, for research purposes only or no derivatives), thus discouraging prospective users, hampering new research and development and leading to repetition in resources creation.

Apart from the above attested challenges, one of the main reasons for the disadvantaged position of Greek is that LT is not included in the overall language policy of Greece and Cyprus, while the recognition of the significance of language-centric AI is still lacking. As a result, a long-term coordinated plan to support LT development in either country is still missing. Sporadic efforts, self-funded or partially supported within programmes in the wider IT or AI areas, have indeed yielded results, but they are not adequate to boost up Greek LT to a state-of-the-art level, nor to help Greek keep pace with language technology developments world-wide. Lack of continuity in research and development funding has been experienced for many years, with short-term projects alternating with longer or shorter periods of drought. While it is important that infrastructural initiatives for language technology have been thriving in Greece, their funding for the future is not secured and their sustainability may be at stake.

In summary, a strategy for keeping Greek up to pace with language technology developments and ensure Greek thrives in the digital sphere should foresee: i) maintenance, extension and sustainability of LT related infrastructures; ii) national and/or European coordinated actions for ensuring access to open high-performance computing infrastructure; iii) coordinated actions for the development of large-scale language resources ready to power large language models supporting a wide range of applications; iv) targeted actions to fill in the observed gaps in speech and multimodal data; v) measures ensuring that the importance of language technology and language-centric AI is appropriately recognised and included in the state policies for language, cultural and technological development; vi) coordinated actions to further enhance digital literacy in the research communities and the society as a whole; vii) coordinated actions to promote the culture of data sharing, including open-source software, involving all stakeholders, the public sector, research and industry.

Acknowledgements

The authors' gratitude goes to Natalia Resende and Bessie Dendrinos. This report has benefited from their insightful comments.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2020.

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.

Wasi Uddin Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. Syntax-augmented multilingual bert for cross-lingual transfer. *arXiv preprint arXiv:2106.02134*, 2021.

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.

Spyros Armosti. The perception of plosive gemination in cypriot greek. *Modern Greek Dialects and Linguistics Theory*, 4(1):39–60, 2009.

Amalia Arvaniti. Cypriot greek. *Journal of the International Phonetic Association*, 29(2):173–178, 1999.

Robert Browning. Lawyer's greek-henrik zilliagus: Zur abundanz der spätgriechischen gebrauchssprache.(societas scientiarum fennica: Commentationes humanarum litterarum, 41.2). pp. 105. helsinki: Suomen tiedeseura, 1967. paper, mk. 8.40. *The Classical Review*, 19(1):67–68, 1969.

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

Athanasia–Lida Dimou, V Pitsikalis, Theodoros Goulas, S Theodorakis, Panagiotis Karioris, M Pissaris, Stavroula-Evita Fotinea, Eleni Efthimiou, and P Maragos. A gsl continuous phrase corpus: Design and acquisition. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC)*, pages 23–26, Istanbul, Turkey, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf.

Eleni Efthimiou, Stavroula-Evita Fotinea, T Hanke, J Glauert, R Bowden, A Braffort, P Maragos, and F Lefebvre-Albaret. Sign language technologies and resources of the dicta-sign project. In *Proc. of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 37–44, Istanbul, Turkey, 2012. URL https://www.researchgate.net/publication/258376046_Sign_Language_technologies_and_resources_of_the_Dicta-Sign_project.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

Maria Gavrilidou. The hellenic national corpus on-line. *Revue belge de philologie et d'histoire*, 80(3): 1003–1015, 2002.

Theodoros Goulas, Dimou Athanasia–Lida, and Eleni Efthimiou. The polytropon parallel corpus. In Mayumi Bono, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, Johanna Mesch, and Yutaka Osugi, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-01-6. URL https://www.sign-lang.uni-hamburg.de/lrec/lrec/pubs/18043.pdf.

Dionysis Goutsos. The corpus of greek texts: a reference corpus for modern greek. *Corpora*, 5(1):29–44, 2010.

Dionysis Goutsos, Georgia Fragaki, Irene Florou, Vasiliki Kakousi, and Paraskevi Savvidou. The diachronic corpus of greek of the 20th century: Design and compilation. In *Proceedings of the 12th international conference on Greek linguistics*, volume 1, pages 369–381, 2017.

Harris Hadjidas and Maria C Vollmer. Multi-cast cypriot greek. *Multi-CAST: Multilingual corpus of annotated spoken texts*, 2015.

Marilena Karyolemou. Η ελληνική γλώσσα στην Κύπρο [the greek language in cyprus]. *Στο Χριστίδης Α.–Φ.(επιμ.) Εγκυκλοπαιδικός Οδηγός για τη γλώσσα. Θεσσαλονίκη: Κέντρο Ελληνικής Γλώσσας*, 2001.

Dimitrios Kasselimis, Maria Varkanitsa, Georgia Angelopoulou, Ioannis Evdokimidis, Dionysis Goutsos, and Constantin Potagas. Word error analysis in aphasia: Introducing the greek aphasia error corpus (graec). *Frontiers in psychology*, 11:1577, 2020.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163, 2014.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117, 2020.

Maria Koutsombogera Maria Gavrilidou, Anastasios Patrikakos, and Stelios Piperidis. Η Ελληνικη Γλωσσα στην Ψηφιακη Εποχη – The Greek Language in the Digital Age. 2012. Available online at http://www.meta-net.eu/whitepapers.

Vassilis Papavassiliou, Prokopidis Prokopis, and Stelios Piperidis. Discovering parallel language resources for training mt engines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/pdf/604.pdf.

Pavlos Pavlou. The semantic adaptation of turkish loan-words in the greek cypriot dialect. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 443–443, 1994.

Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Ion Androutsopoulos, and Constantine D Spyropoulos. Ellogon: A new text engineering platform. *arXiv preprint cs/0205017*, 2002.

Stelios Piperidis. The meta-share language resources sharing infrastructure: Principles, challenges, solutions. In Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Nicoletta Calzolari, and Khalid Choukri, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/1086_Paper.pdf.

Stelios Piperidis, Penny Labropoulou, and Maria Gavriilidou. clarin:el: a language resources documentation, sharing and processing infrastructure [in greek]. In Thanasis Georgakopoulos, Theodossia-Soula Pavlidou, Miltos Pehlivanos, Artemis Alexiadou, Jannis Androutsopoulos, Alexis Kalokairinos, Stavros Skopeteas, and Katerina Stathi, editors, *Proceedings of the 12th International Conference on Greek Linguistics*, volume 2, page 851–869, Berlin, 2017. Edition Romiosini/CeMoG. ISBN 978-3-946142-35-5.

Stelios Piperidis, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou. Managing public sector data for multilingual applications development. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018. URL https://www.aclweb.org/anthology/L18-1205.

Prokopis Prokopidis and Stelios Piperidis. A neural nlp toolkit for greek. In *11th Hellenic Conference on Artificial Intelligence*, pages 125–128, 2020.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. European Language Grid: An Overview. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3359–3373, Marseille, France, 5 2020. European Language Resources Association (ELRA).

Hanna Sababa and Athena Stassopoulou. A classifier to distinguish between cypriot greek and standard modern greek. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 251–255. IEEE, 2018.

Lydia Sciriha. *A question of identity: Language use in Cyprus*. Intercollege Press, 1996.

Marina Terkourafi. Perceptions of difference in the greek spherethe case of cyprus. *Journal of Greek Linguistics*, 8(1):60–96, 2007.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer, 2012.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

Spyridoula Varlokosta, Spyridoula Stamouli, Athanassios Karasimos, Georgios Markopoulos, Maria Kakavoulia, Michaela Nerantzini, Aikaterini Pantoula, Valantis Fyndanis, Alexandra Economou, and Athanassios Protopapas. A greek corpus of aphasic discourse: collection, transcription, and annotation specifications. In *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), Monday 23rd of May 2016*, number 128. Linköping University Electronic Press, 2016.

Μαριάννα Κατσογιάννου and Χαράλαμπος Χριστοδούλου. cyslang: Το λεξικό της κυπριακής αργκό [cyslang: The dictionary of cypriot slang language]. *Modern Greek Dialects and Linguistics Theory*, 7 (1):106–114, 2019.