EUROPEAN
LANGUAGE
EQUALITY

D1.13

Report on the Finnish Language

Authors	Krister Lindén, Wilhelmina Dyster
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	D1.13
Deliverable title	Report on the Finnish Language
Type	Report
Number of pages	24
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe’s Languages in 2020/2021
Authors	Krister Lindén, Wilhelmina Dyster
Reviewers	Andy Way, Sabine Kirchmeier
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
ASR	Automatic Speech Recognition
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DLE	Digital Language Equality
DSM	Digital Single Market
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
EU	European Union
GDPR	General Data protection Regulation
HPC	High-Performance Computing
LR	Language Resource/Resources
LT	Language Technology/Technologies
LUMI	Large Unified Modern Infrastructure
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
SME	Small and Medium-sized Enterprise

Abstract

Language-centric AI is already ubiquitous and language technology is in its intrinsic core. As was stated in the report *The Finnish Language in the Digital Age* (Koskenniemi et al., 2012): “If there is adequate language technology available, it will be able to ensure the survival of languages with small populations of speakers.”

During the last ten years, digitalisation has changed the way we communicate and interact in the world creating an increasing demand for language-based AI services. New skills are needed to be able to cope in the digital world, so digital education and media awareness are now taught in elementary schools. Digital skills are considered new citizen skills.

To provide language-based services to an increasing number of users, we need applications that are built on AI, as well as to provide routine services to special groups and to meet accessibility requirements. The still small number of existing applications and services is partly due to the lack of language resources. Also, the small size of the Finnish market area has affected this when large corporations have primarily focused on English with only some support for Finnish in high-demand products in the Finnish market.

In the field of language technology, the Finnish language is still only moderately equipped with products, technologies and resources. There are applications and tools for speech synthesis, speech recognition, information retrieval, spelling correction and grammar checking. There are also a few applications for automatically translating language. The situation has improved during the last 10 years, but still support for automated translation leaves room for ample improvement and the general support for spoken language is modest in industry applications although some recent research results are encouraging.

Information and communication technologies are preparing for the next revolution using neural networks. With mobile devices and cloud-computing, the next generation of technology will feature software that understands not just spoken or written words and sentences, but supports users far better because it speaks, knows and understands their language.

Forerunners of such developments are the free online services such as Google Translate that translates more than 100 languages including Finnish at a moderately correct level and Apple’s mobile assistant Siri which can react to voice commands and answer questions in more than 35 language varieties. This is a doubling of the coverage since ten years ago.

However, the vision is still that the next generation of information technology will master human language to such an extent that human users will be able to communicate using web services and technology in their own language. Devices will be able to automatically find the most important news and information from the world’s digital knowledge store in reaction to easy-to-use voice commands.

Language-enabled technology will be able to translate automatically or assist interpreters; summarise conversations and documents; and support users in learning scenarios. For example, it will help immigrants and skilled labour to learn the Finnish language and to integrate more fully into the country’s culture as well as enable telecommuting to participate as distance workers in an integrated EU labor market.

The next generation of information and communication technologies will also enable industrial and service robots. The technology must move to modeling language in an all-encompassing way to understand the essence of questions and generate rich and relevant answers. This may require giga-scale data sets which may even be difficult to achieve in small languages as well as in specialised domains of any language, which points to a need for technology leveraging cross-language and cross-domain language-centric AI benefiting from local adaptation with specialised data sets.

In this report, we take stock of the existing resources for Finnish and try to identify some remaining gaps and shortcomings.

Tiivistelmä

Kielikeskeinen tekoäly on jo läsnä kaikkialla, ja keskeisessä osassa sen ytimessä on kieliteknologia. Raportissa *Suomen kieli digitaalisella aikakaudella* (Koskenniemi et al., 2012) todettiin: “Myös pienet kielet selviytyvät varmemmin, jos niille on saatavilla sopivia kieliteknologisia välineitä, jotka tukevat kielen tietokonevälitteistä käyttöä”. Kymmenen viime vuoden aikana digitalisaatio on muuttanut tapaa, jolla viestimme ja olemme vuorovaikutuksessa toistemme kanssa. Tämä on luonut kasvavaa kysyntää myös kieliperusteisille tekoälypalveluille. Uusia taitoja tarvitaan, jotta pärjäämme digitaalisessa maailmassa, joten digitaaliset taidot ja mediatietoisuus ovat nyt mukana peruskoulujen opetussuunnitelmassa. Digitaitoja pidetään uusina kansalaistaitoina.

Jotta kieliperusteisia palveluja voitaisiin tarjota yhä useammille käyttäjille, tarvitaan tekoälyyn perustuvia sovelluksia ja esteettömyysvaatimusten täyttämistä kaikille sekä tavanomaisten palvelujen tuottamista myös erityisryhmille. Olemassa olevien suomenkielisten sovellusten ja palvelujen vähäinen määrä johtuu osittain kieliresurssien puutteesta. Lisäksi Suomen markkina-alueen pienellä koolla on vaikutusta, sillä suuret yritykset keskittyvät ensisijaisesti englanninkielisiin ratkaisuihin, ja vain kaikista kysytyimmille tuotteille on markkina-alueella tarjolla suomenkielistä tukea.

Kieliteknologian alalla suomi on edelleen vain kohtalaisesti tuettu kieli erilaisten tuotteiden, teknologioiden ja resurssien osalta. Suomen kielelle on saatavilla sovelluksia puhesynteesiin, puheentunnistukseen, tiedonhakuun, oikolukuun, kieliopin tarkistukseen ja automaattiseen kääntämiseen. Tilanne on kohentunut 10 viime vuoden aikana, mutta automaattisen kääntämisen tuessa on vielä huomattavasti parantamisen tarvetta. Lisäksi yleinen puhutun kielen tuki on kaupallisissa sovelluksissa vielä vaatimatonta, vaikkakin eräät viimeaikaiset tulokset tutkimuskentältä ovat olleet rohkaisevia.

Tieto- ja viestintäteknologiassa valmistaudutaan seuraavaan vallankumoukseen, jossa hyödynnetään neuroverkkoja. Mobiililaitteiden ja pilvilaskennan ansiosta seuraavan sukupolven teknologia tarjoaa meille ohjelmistoja, jotka pystyvät tukemaan käyttäjää aiempaa kokonaisvaltaisemmin tuottamalla, puhumalla ja ymmärtämällä käyttäjän omaa kieltä. Tällaisen kehityssuunnan edelläkävijöitä ovat ilmainen Google Translate -verkkopalvelu, joka tekee kohtalaisella tasolla käännöksiä yli 100 kielen välillä, suomi mukaan luettuna, ja Applen mobiiliavustaja Siri, joka reagoi äänikomentoihin ja vastaa kysymyksiin yli 35 eri kielellä. Sirin kielitarjonnan kattavuus on kaksinkertaistunut kymmenessä vuodessa.

Visiona on edelleen, että seuraavan sukupolven tietotekniikka hallitsee ihmiskielen siinä määrin, että ihmiset pystyvät kommunikoimaan verkkopalvelujen ja teknologian välityksellä omalla kielellään. Laitteet pystyvät löytämään automaattisesti tärkeimmät uutiset ja tiedot maailman digitaalisesta tietovarastosta reagoimalla helppokäyttöisiin äänikäyttöliittymiin. Lisäksi kieleen perustuva teknologia pystyy tekemään automaattisia käännöksiä tai avustamaan tulkkeja, tekemään yhteenvetoja keskusteluista ja asiakirjoista sekä tukemaan käyttäjiä oppimistilanteissa. Teknologia voisi esimerkiksi auttaa maahanmuuttajia ja ammattitaitoista työvoimaa suomen kielen oppimisessa ja integroitumisessa paremmin maan kulttuuriin sekä mahdollistamaan etätyötä ETA:n yhteisellä työmarkkina-alueella.

Seuraavan sukupolven tieto- ja viestintäteknologia mahdollistaa myös teollisuus- ja palvelurobotit. Jotta niille tulisi kyky ymmärtää kysymyksiä ja antaa monipuolisia, merkityksellisiä vastauksia, pitäisi siirtyä kielen kaikenkattavaan mallintamiseen. Tähän tarvitaan giga-luokan tietoaineistoja, joita on vaikea saada pienille kielille tai erityisalalle. Tarvitaan kieliteknologiaa, jolla on perustoimintakyky kielestä tai alasta riippumatta ja jota pystytään paikallisesti sopeuttamaan eri variantteihin tai tarkoituksiin pienten tai keskikokoisten aineistojen avulla.

Raportissa kartoitetaan, millaisia kieliteknologisia resursseja suomen kielelle on olemassa, ja pyritään tunnistamaan puutteita niin kattavuuden, laadun kuin saatavuuden suhteen.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at European and national/regional levels, language communities, journalists, etc. and it seeks 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 – identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

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

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

2 The Finnish Language in the Digital Age

2.1 General Facts

Finnish is the native language of approximately 4.9 million people living in Finland and the second language of 0.5 million Finns. Finnish is also spoken in Sweden, Estonia, Russia, the United States and Australia. This section is an updated version of *Finnish in the Digital Age* (Koskenniemi et al., 2012).

Finnish is one of the official languages in the European Union. The Finnish constitutional law and language law define Finnish and Swedish as the national languages of Finland. In addition, Finnish is an official minority language in Sweden, in 2020 in 66 municipalities, mainly in Northern and Central Sweden. Besides Finnish and Swedish, three Sámi languages (Northern Sámi, Inari Sámi and Skolt Sámi), Romany, the Karelian language and two different sign languages have long been used in Finland. From the 19th century onwards also Russian- and Tatar-speaking people have been living in Finland. Since the end of the 1970's immigrants have arrived from Europe, Asia and Africa, and the amount of immigrant languages is somewhere around 150, with the major ones being Russian, Estonian, Arabic, English and Somali.

The Finnish literary language has a relatively short history. It has been used in religious literature and the church since the 16th century, and laws have been written in Finnish since the 18th century. Up until the 19th century, Swedish was used in administration, education and literature. The foundation of contemporary Finnish was laid during the 19th century when Finnish became a sovereign language in all societal activity.

Dialects are divided into two categories: the Western and the Eastern dialects. The Western dialects include the South-West dialects, Southern-Western middle dialects, Tavastian dialects, Southern Ostrobothnian dialect, Central and Northern Ostrobothnian dialects and the Peräpohjola dialects. The Eastern dialects include the Savonian dialects and the South-Eastern dialects. The difference between the Eastern and Western dialects is mostly in the pronunciation and word forms (*meijän*, *männä* in the East while *meirän*, *mennä* in the West)

[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.

and partly in the vocabulary (*vasta* in the East, *vihta* in the West.) The differences between dialects are clear, and speakers from different areas can be identified by their intonation. However, the differences are minor enough to allow speakers of different dialects to understand each other. Urbanisation and other changes in society have softened the dialects and smoothed out the most narrow and distinctive features.

2.2 Finnish in the Digital Sphere

Finnish is used widely and actively on the Internet and on social media. Almost all Finnish households (96%) have access to the Internet. Around half of all households have both mobile and broadband access, and 43% of households rely solely on mobile technology, which means that the development of 5G/6G network technology is essential for a large number of Finns to stay connected to the digital sphere.[2] Traficom, the Finnish Transport and Communications Agency, reported in November 2020 that the total number of registered FI-domains had reached 500,000. The previous milestone of 250,000 registrations was reached by the end of 2010, suggesting a growth of 100% in only 10 years. The global COVID-19 pandemic has launched a rapid need for functional e-commerce sites and digital services, and this change is also visible in the statistics as a strong peak in the number of new recently registered FI-domains.[3]

3 What is Language Technology?

Natural language[4] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

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

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

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various

[2] https://tieto.traficom.fi/fi/tilastot/laajakaistayhteyksien-levinneisyys-kotitalouksissa

[3] https://www.epressi.com/tiedotteet/teknologia/digitaalisten-palveluiden-rooli-korostunut-poikkeuksellisena-aikana-fi-verkkotunnusten-maarassa-saavutettiin-puolen-miljoonan-rajapyykki.html

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

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 a few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance, for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[5]

4 Language Technology for Finnish

The development of Finnish language data and tools has progressed steadily over the past 30 years. Since 1995, the *Language Bank of Finland*[6] and since 2015 CLARIN and FIN-CLARIN have offered a wide variety of text and speech corpora and tools for studying them. Today, a large number of fundamental tools and datasets are available for Finnish. Below we present some relevant resources in the different domains of LT. For additional resources, see *META-SHARE Finland*.[7] Nevertheless, more work remains to be done in building domain-specific corpora and speech processing components.

4.1 Language Data

Monolingual text corpora

There are several large monolingual corpora, which contain contemporary language use. *The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland* (National Library of Finland, 2011a) is a corpus of Finnish newspapers and magazines dating from 1820. Besides historical entries, it contains entries until the 1940s. Another recently updated corpus of the same domain is the *Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s* (University of Helsinki, 2019). Online discussions are featured in the *The Suomi24 Corpus* (City Digital Group, 2021), which covers all the discussion forums of the Suomi24 online social networking website from 2001 to 2020. This corpus is licensed for academic use. *The Finnish OpenSubtitles (OPUS)* (Huovilainen, 2018a) corpus contains Finnish subtitles for movies and TV-series. All corpora have been tokenised and annotated with morpho-syntactic analysis produced with the *Turku Dependency Parser*.[8]

Overall, general domain data seems to be prevalent, e. g. data collected from discussion forums or using web crawls. In addition, news texts, legislative texts and parliamentary speech are well-represented domains. The Language Bank of Finland has the expertise to handle sensitive health data, but health domain corpora are still scarce.

[5] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18,4% from 2020 to 2028 (https://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).

[6] https://kielipankki.fi

[7] https://metashare.csc.fi

[8] https://github.com/TurkuNLP/Finnish-dep-parser

Bi- and multi-lingual text corpora

Some of the largest bi-lingual corpora include *The Newspaper and Periodical Corpus of the National Library of Finland* (National Library of Finland, 2011b) and *KOTUS Finnish-Swedish Parallel Corpus* (Institute for the Languages of Finland, 2015), which are both for the language pair Finnish-Swedish. For the pair Finnish-English, some of the largest corpora have been built through web-crawl data collection efforts, e. g. *ParaCrawl*[9] and *Finnish web corpus fiWaC* (Ljubešić et al., 2016), which was built by crawling the .fi top-level domain in 2015 for both Finnish and English documents.

Multimodal corpora (audio, video)

A Multimodal Corpus of Tourist Brochures Produced by the City of Helsinki, Finland (1967-2008) (Hiippala, 2015) is fully annotated using XML schema provided for the Genre and Multimodality (GeM) model (Bateman, 2008). There are also several different corpora for Finnish Sign Language, which typically contain both video and audio. One of these is *Kipo Corpus* (Kuurojen Liitto ry, 2015), which contains the language policy programme of the sign languages of Finland in Finnish and Finnish Sign Language. *Hundred Finnish Linguistic Life Stories*[10] is a multilingual corpus, which contains images and annotated interviews that are synchronised with audio and video. *The Yle MeMAD Media Corpus*[11] contains selected TV programmes and videos with their descriptive metadata and subtitles from the archives of The Finnish Broadcasting Company (Yle), from 1966 to 2018. The main audio and subtitle languages are Finnish and Swedish with some content in English. Corpus use outside the MeMAD project needs to be licensed separately. However, in late 2021, Yle has released three datasets with an experimental licence for a limited amount of time to support the development of language and media-related technologies.[12]

The largest corpora containing modern Finnish speech are *Aalto Finnish Parliament ASR Corpus 2008-2020* (3000 h) (Aalto University, Department of Signal Processing and Acoustics, 2022) and the *Donate Speech Corpus* (4000 h).[13] The latter corpus was collected in a campaign that started in the summer of 2020, with the goal to gather ordinary, casual Finnish speech that can be used for studying language as well as for developing technology and services that can be readily used in Finnish. The Donate Speech Corpus is licensed to allow for both academic and commercial use of the material under given terms.

Lexical/conceptual resources

The Institute for the Languages of Finland has comprehensive collections of lexical corpora. *Word Collections of Modern Finnish*[14] is a corpus of over 5,5 million entries of Finnish words. Each entry contains the reference, its passage and information on the original context. The corpus has been collected mostly from literature, newspapers and magazines. Entries from 1984 and later are in digital form and require a permission to use. *Dictionary of Contemporary Finnish*[15] is a dictionary of standard Finnish, which contains over 100,000 lemmas and provides information on the meanings, usage and nuances of style of contemporary Finnish words, as well as their inflection and spelling. *The dictionary of Finnish dialects*[16] is licensed

[9] https://paracrawl.eu
[10] http://urn.fi/urn:nbn:fi:lb-2019092003
[11] https://metashare.csc.fi/repository/browse/the-yle-memad-media-corpus/bffe36ae94d211e98e73005056be118ea48a924c81974d8893a
[12] https://developer.yle.fi/en/data/avdata/index.html
[13] http://urn.fi/urn:nbn:fi:lb-2020090321
[14] http://urn.fi/urn:nbn:fi:lb-20140730187
[15] https://www.kielitoimistonsanakirja.fi
[16] https://kaino.kotus.fi/sms/

under the CC-BY 4.0 licence and it is accessible online. The development of the dictionary still continues and around 6,000 new entries are yearly added to the resource. Once finished, the dictionary will contain around 350,000 word entries.

Some of the large lexical/conceptual corpora for the Finnish language have been collected through crowd-sourced projects or in projects that have been co-funded by the EU. One example of the latter is the *Finnish WordNet* (University of Helsinki, 2010), which contains words grouped by meaning into synonym groups representing concepts and they are linked to each other creating a semantic network. The *FinnWordNet* is licensed under the CC-BY 3.0 licence, and it can be used in LT research and applications or as an electronic thesaurus. However, it is not currently being actively developed and the most recent version was released in 2012.

ConceptNet[17] is a freely-available, multilingual semantic network, which contains a Finnish vocabulary of the size of 380,000 terms. It is used to create word embeddings that are aligned across languages and designed to avoid representing harmful stereotypes. *ConceptNet* originated from the crowd-sourcing project *Open Mind Common Sense*, launched in 1999 at the MIT Media Lab.

The Helsinki Term Bank for the Arts and Sciences (HTB)[18] is a multidisciplinary project which aims to gather a permanent terminological database for all fields of research in Finland. The working method is a type of limited crowd-sourcing as the terminology will be gathered among expert groups in different fields of research.

Models and grammars

- *Psycholinguistic Descriptives* (Huovilainen, 2018b) comprises a dataset with frequency information for words, lemmas, syllables and letter n-grams for Finnish. The dataset is based on six large corpora from sources such as magazines, newspapers, movie and tv-series subtitles, encyclopedia topics and Internet discussions that together comprise of 2.5 billion words.
- *Crúbadán language data for Finnish* (Scannell, 2011) is a dataset containing word and character n-gram frequencies. This language resource was created from web-crawled corpora and all the data are freely accessible, and made available under a CC-BY license.
- *The Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland* (National Library of Finland, 2014) is a resource that contains sets of word unigrams, bigrams and trigrams extracted by the University of Helsinki from the source data. The n-grams have been computed across sentence boundaries for each decade (from the 1820s to the 2000s) as well as for the entire corpus, and the frequencies derived from it are available under the CC BY 4.0 license.
- *Comprehensive Grammar of Finnish*[19] was published in 2004 by the Finnish Literature Society. The online version, unofficially VISK, includes this grammar, its terminology published in 2005 with definitions, and a number of features to facilitate information retrieval.
- *FinEst BERT*[20] offers a multilingual model trained from scratch, covering three languages: Finnish, Estonian, and English. This model can be used for various NLP classification tasks, supporting both monolingual and multilingual/crosslingual (knowledge transfer) tasks.

[17] https://conceptnet.io
[18] https://tieteentermipankki.fi
[19] https://kaino.kotus.fi/visk/
[20] http://urn.fi/urn:nbn:fi:lb-2020061201

- *FinBERT*[21] is a version of Google's BERT deep transfer learning model for Finnish, developed by the TurkuNLP Group. FinBERT has been pre-trained for 1 million steps on over 3 billion tokens of Finnish text drawn from news, online discussion, and internet crawls.

4.2 Language Technologies and Tools

Text Analysis

- *The Helsinki Finite-State Transducer (HFST)*[22] software is intended for the implementation of morphological analyzers and other tools which are based on weighted and unweighted finite-state transducer technology.
- *Finnish dependency parser developed by TurkuNLP*[23] is an open source dependency parsing pipeline for analyzing Finnish text.

Speech Processing

- *Aalto University Automatic Speech Recognition System* Aalto-ASR[24] is a toolkit that provides functionalities for automatic speech recognition from audio files and for automatic forced alignment of text and speech. The current version includes models for recognising Finnish speech and for aligning speech recordings with transcripts in Finnish, Swedish or Northern Sami.

Translation Technologies

- *OPUS-MT* (Tiedemann and Thottingal, 2020) is a project that focuses on the development of free resources and tools for machine translation. The current status is a repository of over 1,000 pre-trained neural machine translation models that are ready to be launched in on-line translation services.

Information Extraction and Information Retrieval

- *Finto AI*[25] is a service for automated subject indexing, which can be used to suggest subjects for texts in Finnish, Swedish and English. Finto AI is based on Annif and it currently gives suggestions based on concepts of the General Finnish Ontology YSO.

Language Generation and Summarisation

- The EMBEDDIA Media Assistant (EMA)[26] is a collection of AI tools for the media sector and text-based industry, supporting a range of tasks and languages. A special focus is on less-resourced European languages, including Finnish.
- The NewsEye[27] project has developed a set of tools and methods that will improve users' capability to access, analyse and use the content in the digital Libraries of historical

[21] https://github.com/TurkuNLP/FinBERT
[22] https://hfst.github.io
[23] https://github.com/TurkuNLP/Finnish-dep-parser
[24] http://urn.fi/urn:nbn:fi:lb-2021082323
[25] https://ai.finto.fi
[26] https://embeddia.texta.ee
[27] https://www.newseye.eu

newspapers. The tools include dynamic text analysis to automatically find topics or viewpoints in the corpus being studied.

Human-Computer Interaction

- Wavelet-based embedding models for speech synthesis for Finnish have been developed at the University of Helsinki.

Some examples from the health and social media domains:

There are only a few examples of health domain corpora for the Finnish language: Multilingual *European Medicines Agency corpus (EMEA)* and monolingual *Medicine Radar (Lääketutka)*,[28] which organises Suomi24 online discussions (2001-2016) where people describe their drug use and symptoms.

The social media domain is covered by the *Suomi24* (City Digital Group, 2021) and *Ylilauta* (Ylilauta, 2015) online discussions corpora, which have content that can be used for training systems to recognise hate speech and propaganda.

The multilingual *Mega-Cov 0.2 corpus* (Abdul-Mageed et al., 2021), described as a 'Billion-scale dataset from Twitter for studying COVID-19 (2007-2020)', probably also contains material suitable for training fake news detection.

4.3 Projects, Initiatives, Stakeholders

In October 2017, the Finnish Ministry of Economic Affairs and Employment published its national AI strategy entitled Finland's age of artificial intelligence[29] (Finland, 2017). This report fits under the umbrella of a broader Artificial Intelligence Programme in Finland (also labelled as **AI Finland**) with a view to establishing AI and robotics as the cornerstones of success for Finnish companies.

In November 2019, VAKE (currently the Climate Fund) published a **report on language-centric artificial intelligence development in Finland** (Jauhiainen et al., 2019) pointing to neural networks suitable for deep learning as well as more traditional methods for machine learning. The report specified the next phase of the language-centric artificial intelligence development program and collected topics in need of interventions.

In February 2020, the Ministry of Finance launched the AuroraAI[30] programme. The task in AuroraAI is to develop an operating model for arranging public administration activities to support people in different life situations and events so that services provided by organisations function seamlessly between service providers in different sectors. Continuing until the end of 2022, the programme lays the foundation for using artificial intelligence to bring services and people together in a better way.

In November 2020, Finland launched an updated national AI strategy. **The Artificial Intelligence 4.0 Programme** promotes the development and introduction of AI and other digital technologies in companies, with a special focus on SMEs. In the first interim report,[31] published in April 2021, the programme presented a vision for the future of the Finnish manufacturing industry, stating that in 2030 the Finnish manufacturing industry will be clean, efficient and digital. As stated in the report, seamless collaboration between high-speed telecommunications networks, cloud computing and AI are central to digital transformation.

[28] https://laaketutka.fi

[29] http://urn.fi/URN:ISBN:978-952-327-290-3

[30] https://valtioneuvosto.fi/en/-/10623/the-auroraai-national-artificial-intelligence-programme-begins-with-the-aim-of-using-artificial-intelligence-to-bring-people-and-services-together-in-a-better-way

[31] http://urn.fi/URN:ISBN:978-952-327-643-7

Finland's AI 4.0 Programme includes the following aims:

- strengthen digitalisation and economic growth in Finland
- encourage cooperation between different sectors, increase investments in digitalisation and improve digital skills in SMEs[32]
- contribute to the recovery of companies and the economy from the coronavirus pandemic.

There are several national research communities that support AI research, but only a handful that are explicitly related to language or LT. Some of the most important projects and applications in the field in the last five years include:

- The Finnish National Broadcasting Company (Yle) in collaboration with the University of Helsinki and the State Development Company VAKE launched a campaign in 2020 Donate Your Speech (Lahjoita puhetta) to collect spoken Finnish from all around the country so algorithms could be taught to understand and recognise different Finnish dialects.
- As for other Finnish universities, Aalto University has established the Aalto Speech Recognition Group, whose research projects have produced several open source speech and language modeling tools, which have been used by multiple companies.
- In 2019, one of the outcomes of the Finnish Presidency of the Council of the European Union was the establishment of the free, widely accessible online university-level AI course called Elements of AI funded by the Ministry of Economic Affairs and Employment and developed by the University of Helsinki and Reaktor Innovations Oy. The course is available in all EU languages. From the inception of the project, over 730,000 students have successfully taken and completed the course, indicating unprecedented levels of global interest in the subject of AI as well as freely accessible education in this domain. Additionally, the University of Helsinki, in collaboration with multiple Finnish and international partners, has developed a 2 ECTS course that involves texts and assignments called Ethics of AI, also available openly online. The Aalto University also offers a Diploma in Artificial Intelligence, a study programme that gives an in-depth understanding of the topic and helps people to understand and apply contemporary AI technologies.

The role of the Language Bank of Finland is to support academic research and to provide some support for industrial use of academic resources which are also available for commercial use.

According to Business Finland, "Initiatives like Finnish AI Accelerator, the Tampere AI Hub and the AI Academy at the University of Turku drive AI commercialisation by effectively transferring knowledge and findings to startups".[33]

Generally, the Finnish market is extremely active in the AI field. According to the 'State of AI in Finland' report by FAIA (2020) "There are over 1250 companies that use different AI applications, of which roughly 750 have developed their own technology."

A rapidly growing startup ecosystem boosts AI/LT development. The FAIA report also states "There are over 400 AI startups in Finland and approximately 50 new companies are established annually."

FAIA curates a landscape of the top AI-first firms in Finland. The newest edition of the landscape was published in June 2020 and consisted of 42 firms.

[32] https://tem.fi/en/-/artificial-intelligence-4.0-programme-to-speed-up-digitalisation-of-business

[33] https://www.businessfinland.fi/en/do-business-with-finland/explore-key-industries/ict-digitalization/ai

CSC – IT Center for Science is tasked with providing one of the three EuroHPC supercomputers, LUMI. The whole system is designed with AI, machine learning and data analytics in mind. LUMI's first pilot phase was concluded by the end of 2021, and LUMI will reach its full capacity in 2022.[34]

5 Cross-Language Comparison

The LT field[35] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded unforeseeable results. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[36] 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.)

[34] https://www.lumi-supercomputer.eu

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

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

5.2 Levels of Technology Support

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

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in $\geq$3% and $<$10% of the ELG resources of the same type
3. **Moderate support**: the language is present in $\geq$10% and $<$30% of the ELG resources of the same type
4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[37]

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[38] 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.[39]

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

[37] 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.

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

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

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

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

5.4 Results and Findings

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

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[40] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

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

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

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

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

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

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

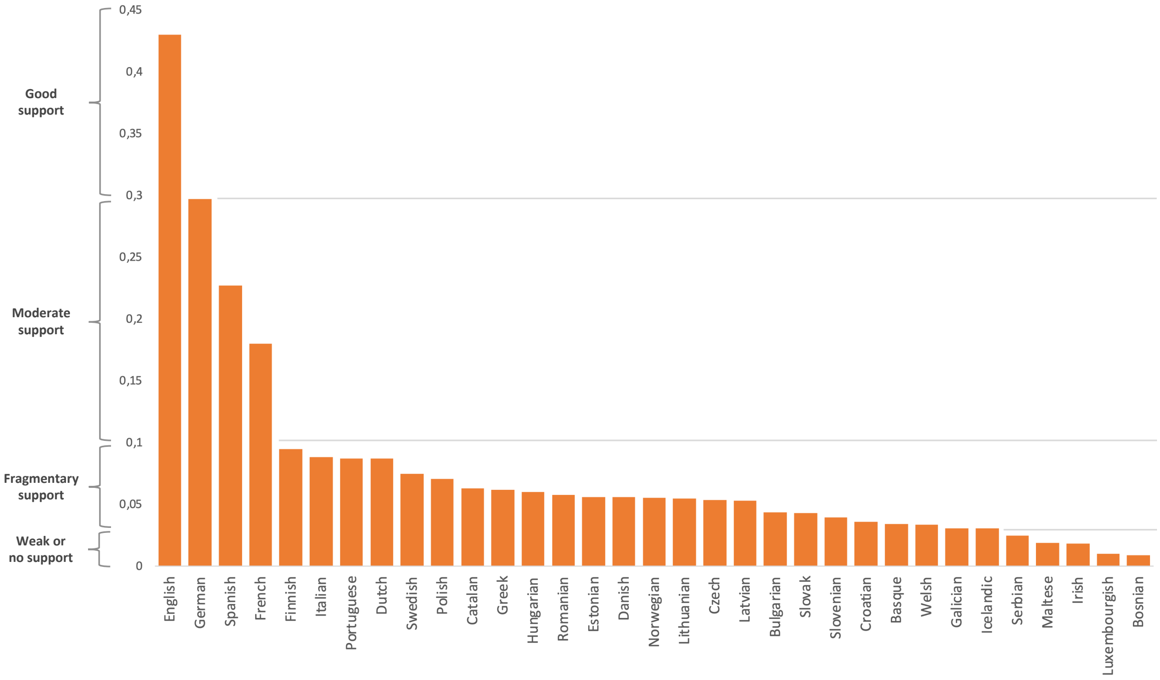

Figure 1: Overall state of technology support for selected European languages (2022)

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

The vision is to let the next generation of information and communication technologies enable language-centric AI services, for which the technology must move to modelling language in an all-encompassing way to understand the essence of questions and generate rich and relevant answers. In this report, we have surveyed the existing resources for Finnish to identify some remaining gaps and shortcomings to fulfill this vision.

As pointed out in the VAKE report (Jauhiainen et al., 2019), we need the availability and accessibility of components for processing speech with open licences (e. g. MIT, CC0 or similar) to create prototypes or develop methods into full-scale production versions in the hands of companies. To facilitate the development, collaboration between different organisations is needed: an ecosystem with a forum or a platform where different-level actors can come together to exchange experiences and seek new projects and collaboration opportunities. There is also a need for expertise in GDPR and legal issues concerning collecting, distributing and using language resources, especially speech resources. This collaboration platform can also become a center for information and education on these topics.

Despite various programmes, initiatives and strategies, there is still a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. We can therefore conclude that there is still a desperate need for a large, coordinated initiative focused on overcoming the differences in language technology readiness for European languages as a whole.

Ten years ago, the situation was as follows: 1) few multi-modal resources and virtually no advanced discourse processing tools available for Finnish, 2) only very few projects working on information retrieval for Finnish, 3) an unclear legal situation restricting the use of digital texts, 4) some specific corpora of high quality, but no large, up-to date resources for product development targeted at the everyday users, 5) no applications based on semantic analysis, 6) in speech technology, the biggest leap forward was in the area of speech recognition, but a breakthrough had not yet been implemented in the commercial sector, and in speech synthesis research, the work was still in the laboratory phase as speech corpora were considered hard to collect.

This is the present situation: 1) There are some multi-modal resources as listed in the report, but still no advanced discourse processing tools for Finnish. 2) Several research projects are working on advanced information retrieval and data mining for Finnish. 3) The legal situation has become clearer with the General Data protection Regulation (GDPR), but we are still waiting for Finland to fully implement the Digital Single Market Directive (DSM). 4) We have some specific corpora of high quality, but the commercial sector in Finland still needs large, up-to date resources for product development targeted at everyday users and technologies to collect specialised data sets. 5) Work on semantics has still not led to significant applications, but this is explored in the context of advanced research projects on information retrieval and extraction. 6) In speech technology, the recent biggest leaps forward have been made using neural network technology. This has also lead to some improvements for the commercial sector offering speech-based services, but speech and video corpora are no longer considered hard to collect with the advent of mobile phones and teleconferencing.

Based on the above, we note that speech corpora and especially resources for spontaneous speech recognition and various genres of speech synthesis are currently being developed. The need for extensive and varied text materials can to some extent be rectified for research purposes through corpus collections of publicly produced language material when properly considering the GDPR and the DSM directive. This will enable the creation of language models based on this material. However, we still need a variety of specialised data sets of language materials for domain-specific purposes to adapt open-source or proprietary software components. Developing dedicated language components from scratch requires giga-scale data sets which may be difficult to compile for small language communities and in specialised domains. This points to a need for a general-purpose language-centric AI which can leverage cross-language and cross-domain resources and benefit from adaptation to local language varieties and specialised domains with small or medium-sized data sets.

References

Aalto University, Department of Signal Processing and Acoustics. Aalto Finnish Parliament ASR Corpus 2008-2020, 2022. URL http://urn.fi/urn:nbn:fi:lb-2021051903.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Dinesh Pabbi, Kunal Verma, and Rannie Lin. Mega-cov: A billion-scale dataset of 100+ languages for covid-19, 2021.

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

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

John Bateman. *Multimodality and Genre, A Foundation for the Systematic Analysis of Multimodal Documents*. Springer, 04 2008. ISBN 978-1-349-28079-7. doi: 10.1057/9780230582323.

Noam Chomsky. *Syntactic Structures*. Mouton and Co., The Hague, 1957.

City Digital Group. The Suomi 24 Corpus 2001-2020, VRT version, 2021. URL http://urn.fi/urn:nbn:fi:lb-2021101527.

Tuomo Hiippala. A Multimodal Corpus of Tourist Brochures Produced by the City of Helsinki, Finland (1967-2008), 2015. URL http://urn.fi/urn:nbn:fi:lb-201411281.

Tatu Huovilainen. Finnish OpenSubtitles 2017, Kielipankki Korp Version, 2018a. URL http://urn.fi/urn:nbn:fi:lb-2018060403.

Tatu Huovilainen. Psycholinguistic Descriptives, 2018b. URL http://urn.fi/urn:nbn:fi:lb-2018081601.

Institute for the Languages of Finland. The Helsinki Korp Versio of the KOTUS Finnish-Swedish Parallel Corpus, 2015. URL http://urn.fi/urn:nbn:fi:lb-2016042704.

Tommi Jauhiainen, Mietta Lennes, Terhi Marttila, et al. Suomenkielisen tekoälyn kehittämisohjelma–esiselvitys, 2019.

Kimmo Koskenniemi, Krister Lindén, Lauri Carlson, Martti Vainio, Antti Arppe, Mietta Lennes, Hanna Westerlund, Mirka Hyvärinen, Imre Bartis, and Pirkko Nuolijärvi. *THE FINNISH LANGUAGE IN THE DIGITAL AGE*. Springer, 2012.

Kuurojen Liitto ry. Kipo-korpus (Suomen viittomakielten kielipoliittinen ohjelma 2010), ladattava versio 2, 2015. URL http://urn.fi/urn:nbn:fi:lb-2020112921.

Nikola Ljubešić, Tommi Pirinen, and Antonio Toral. Finnish web corpus fiWaC 1.0, 2016. URL http://hdl.handle.net/11356/1074. Slovenian language resource repository CLARIN.SI.

National Library of Finland. The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version, 2011a. URL http://urn.fi/urn:nbn:fi:lb-2016050302.

National Library of Finland. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874), 2011b. URL http://urn.fi/urn:nbn:fi:lb-2015051201.

National Library of Finland. The Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland, 2014. URL http://urn.fi/urn:nbn:fi:lb-2014073038.

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.

Kevin P Scannell. Statistical unicodification of african languages. *Language resources and evaluation*, 45(3):375–386, 2011.

Jörg Tiedemann and Santhosh Thottingal. Opus-mt – building open translation services for the world. In André Martins [et al.], editor, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Switzerland, November 2020. European Association for Machine Translation. URL https://eamt2020.inesc-id.pt/. Annual Conference of the European Association for Machine Translation , EAMT2020 ; Conference date: 03-11-2020 Through 05-11-2020.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 00264423, 14602113. URL http://www.jstor.org/stable/2251299.

University of Helsinki. The Downloadable Version of the Finnish WordNet, 2010. URL http://urn.fi/urn:nbn:fi:lb-2016042503.

University of Helsinki. Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s (VRT), Version 2, 2019. URL http://urn.fi/urn:nbn:fi:lb-201908191.

Ylilauta. Ylilauta Corpus, 2015. URL http://urn.fi/urn:nbn:fi:lb-2015031802.