EUROPEAN LANGUAGE EQUALITY

# D1.12

# Report on the Estonian Language

| | |
|---|---|
| Author | Kadri Muischnek |
| 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 | D1.12 |
| Deliverable title | Report on the Estonian Language |
| Type | Report |
| Number of pages | 20 |
| 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 |
| Author | Kadri Muischnek |
| Reviewers | Andy Way, Annika Grützner-Zahn |
| Editors | Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne |
| EC project officers | Susan Fraser, Miklos Druskoczi |
| Contact | European Language Equality (ELE)<br>ADAPT Centre, Dublin City University<br>Glasnevin, Dublin 9, Ireland<br>Prof. Dr. Andy Way – andy.way@adaptcentre.ie<br>European Language Equality (ELE)<br>DFKI GmbH<br>Alt-Moabit 91c, 10559 Berlin, Germany<br>Prof. Dr. Georg Rehm – georg.rehm@dfki.de<br>http://www.european-language-equality.eu |
| |  |

# 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

**1 Introduction** **2**

**2 The Estonian Language in the Digital Age** **2**
2.1 General Facts . . . . . 2
2.2 Estonian in the Digital Sphere . . . . . 3

**3 What is Language Technology?** **3**

**4 Language Technology for Estonian** **5**
4.1 Language Data . . . . . 5
4.2 Language Technologies and Tools . . . . . 7
4.3 Projects, Initiatives, Stakeholders . . . . . 8

**5 Cross-Language Comparison** **9**
5.1 Dimensions and Types of Resources . . . . . 9
5.2 Levels of Technology Support . . . . . 10
5.3 European Language Grid as Ground Truth . . . . . 10
5.4 Results and Findings . . . . . 11

**6 Summary and Conclusions** **14**

# List of Figures

1 Overall state of technology support for selected European languages (2022) . . 13

# List of Tables

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) . . . . . 12

# List of Acronyms

| | |
|---|---|
| AI | Artificial Intelligence |
| ASR | Automatic Speech Recognition |
| CEF | Connecting Europe Facility |
| CELR | Centre of Estonian Language Resources |
| CL | Computational Linguistics |
| CLARIN | Common Language Resources and Technology Infrastructure |
| DLE | Digital Language Equality |
| DH | Digital Humanities |
| ELE | European Language Equality *(this project)* |
| ELE Programme | European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)* |
| ELG | European Language Grid (EU project, 2019-2022) |
| ELRC | European Language Resource Coordination |
| EU | European Union |
| GPU | Graphics Processing Unit |
| HPC | High-Performance Computing |
| IT | Information Technology |
| LR | Language Resource/Resources |
| LT | Language Technology/Technologies |
| META | Multilingual Europe Technology Alliance |
| META-NET | EU Network of Excellence to foster META |
| ML | Machine Learning |
| MT | Machine Translation |
| NER | Named-Entity Recognition |
| NLG | Natural Language Generation |
| NLP | Natural Language Processing |
| NLTP | National Language Technology Platform |
| UD | Universal Dependencies |

## Abstract

This report gives an overview of the state of the art of the language resources and tools for Estonian and identifies the gaps.

The Estonian language has only around one million speakers and so the market for language technology products for Estonian is also a small one. The main force driving the development of Estonian language technology has been the public sector and so the resources and tools developed by state-funded projects are open source, but tend rather to be prototypes, not finished products. Still, during the last decade the situation has been improving and the private sector has also engaged in creating tools and solutions for Estonian language technology. The last decade has also seen considerable growth and development in language resources for Estonian, but several gaps that were identified by the Meta-Net White Paper ten years ago are still there.

The large monolingual corpora are collected regularly and they are large enough to be used for building massive language models. Estonian is included in the multilingual resources of the EU languages, but there is too few parallel data that is a result of direct translation between other language pairs than English-Estonian.

Estonian has at least a minimal necessary amount of audio resources for Estonian, but more and/or bigger special corpora are needed.

Lexical-conceptual resources of Estonian are mostly lexicons, dictionaries and machine-readable dictionaries, as well as terminological databases. Majority of these resources were originally developed as basis for human-readable dictionaries. A significant exception is the Estonian Wordnet.

There is one full-coverage rule-based computational grammar for Estonian, namely Constraint Grammar. However, there are several massive monolingual models, e. g., EstBERT, Estonian RoBERTa, ELMo, a brand-new GPT2 and also a domain-specific model WIKIBert-et.

Estonian has the basic language technology tools for text analysis, speech recognition and speech synthesis, as well as for machine translation, but lacks resources and tools for computational semantics, Estonian Wordnet being a notable exception.

There are quite good basic text analysis tools – sentence segmentation, tokenisation, morphological analysis, syntactic parsing – for standard written language. As soon as the text deviates from the standard, the quality of the analysis decreases. So the processing of the language used on social media sites is more problematic.

Although the quality of speech processing tools and services is far from the quality of those for English, the situation for Estonian is quite good, at least for “ideal” speech, i. e., while the speaker is speaking Estonian as their first language and has no specific health conditions and there is little background noise. There are also several models for speech synthesis, including a neural network-based one.

The need for LT support has been acknowledged by Estonian government agencies and policy-makers. Since 2006 there has been a series of National Programmes for Language Technology, with the current one in force until the year 2027. LT is also part of Estonia’s strategical plan for AI and of the official Estonian Language Development Plan.


## Eesti keele keeletehnoloogiline tugi

See dokument kirjeldab eesti keele keeletehnoloogilise toe olukorda 2021. aastal teiste Euroopa Liidu keelte olukorra taustal.

*Keeletehnoloogia* on interdistsiplinaarne teaduse ja tehnika valdkond, mis tegeleb inimkeelt töötlevate, analüüsivate, genereerivate ja mõistvate süsteemide uurimise ja väljatööta-

misega. Selle dokumendi 3.õsas kirjeldatakse lühidalt keeletehnoloogia valdkondadi ja tähtsamaid rakendusi.

4.õsas antakse ülevaade eesti keele tähtsamatest keeleressurssidest ning keeletehnoloogilistest tööriistadest. Keeleressursside all mõeldakse üks- ja mitmekeelseid tekstikogusid ehk keelekorpusi, multimodaalseid kogusid, nt kõnekorpusi, leksikone ja tesaurusi ning formaalseid grammatikaid ja keelemudeleid.

Keeletehnoloogiliste tööriistade all käsitletakse esiteks üldisi tekstianalüüsi vahendeid – lausestamise, sõnestamise, morfoloogilise ja süntaktilise analüüsi tööriistu. Antakse ülevaade ka eesti keele kõnetehnoloogia, masintõlke, infoeralduse ja arvutiga loomulikus keeles suhtlemise vahenditest.

Üldiselt võib öelda, et eesti keele keeletehnoloogilise toe olukord on rahuldav ja nagu võib näha 5.õsas esitatud joonisel, on see teiste EL riikides räägitavate keeltega võrreldes keskmisel tasemel. Osas 5. võrreldaksegi Euroopa Liidu keelte tehnoloogilise toe tasemeid ning esitatakse selle võrdluse lähtekohad.

# 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 – 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 and experts in more than 30 European languages have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed by 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 Estonian Language in the Digital Age

## 2.1 General Facts

Differently from most languages spoken in Europe, Estonian is not an Indo-European language, but belongs to the Balto-Finnic group of the Finno-Ugric languages. Its closest relatives are the near-extinct Livonian and Votian languages. Other Balto-Finnic languages include Finnish, Karelian, Ingrian, Veps and Ludic.

As there is no clear-cut distinction between a language and a dialect, some authors describe Võro and Seto as languages, and others as dialects of Estonian. If regarded as languages, they are the closest relatives of Estonian.

Typologically, Estonian represents a transitional form from an agglutinating to a fusional language. The characteristic features of Estonian include the accent on the first syllable, a

[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.

high frequency of vowels as opposed to consonants, three different lengths of vowels and consonants, the lack of grammatical gender and articles, and a basic vocabulary different from that of the Indo-European languages.

Estonian has a rich morphological system: nominals inflect for case and number, and verbs for person, number, tense, mood and voice. Compounding is relatively free and productive in Estonian, and compounds are written as one word-form. Derivation is another productive device for forming new lexical items. The word order of Estonian is rather free and mostly governed by information structure. The most important rule is V2: the verb occupies the second position in the clause. For a thorough linguistic description of Estonian, the reader should refer to Erelt (2003).

Estonian is written using a supplemented Latin alphabet; in addition to ASCII characters, it also includes the letters Ä, Ö, Ü, Õ, Š and Ž. Keyboards and laptops commercially available in Estonia are customised for writing in Estonian.

Estonian is the official language of the Republic of Estonia. The overall population of Estonian is ca. 1.3 million and according to Statistics Estonia,[2] there were ca. 915,000 people in Estonia identifying themselves as Estonians in the year 2018. For all of these people, Estonian – one of the official languages of the European Union – is most likely their first language.

The Estonian language is used in all spheres of life: as an administrative language, in media, in education and higher education, science and in cultural spheres, e. g., fiction and theatres.

However, there are some concerns regarding the use of Estonian in science and higher education. Obviously, English is the language of international scientific communication and Estonian researchers are part of this international community. But as a side-effect of this, publishing in Estonian is sometimes regarded as somewhat pointless, which may result in the impoverishment of scientific terminology in Estonian.

Furthermore, some curricula at master's and doctoral levels are offered in English only, which also contributes to the impoverishment of Estonian scientific terminology and Estonian-language scientific writing and communication skills in general. In IT curricula, this trend is even stronger.

### 2.2 Estonian in the Digital Sphere

The Estonian population has good access to internet and digital services: 92% of Estonian households have an internet connection at home[3] and a lot of services are available online. For example, one can declare one's income and ask for a tax refund at the website of Estonian Tax and Customs Board, view one's medical data, submit statements of intention via the Patient Portal etc.

According to a recent study, Estonian children spend on average 172 minutes on the internet every day (Smahel et al., 2020).

On the 1st November 2021, the number of websites with Estonia's code as the top-level domain was 146,264.[4]

## 3 What is Language Technology?

Natural language[5] 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

[2] https://andmed.stat.ee/en/stat
[3] https://andmed.stat.ee/en/stat/majandus
[4] https://www.internet.ee
[5] 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).

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

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

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

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been displaced by data-based ones, i. e., systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

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

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

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e., the generation of speech given a piece of text, Automatic Speech Recognition (ASR), 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.[6]

# 4 Language Technology for Estonian

## 4.1 Language Data

Large monolingual Estonian web corpora have been collected regularly (in 2013, 2017, and 2019, with a new corpus to be published during the first half of 2022), and they are large enough to be used for building massive language models.

In specific domains, e g. court decisions or healthcare, large text collections exist, but they can be used only under very strict constraints. The main obstacles are ethical and political considerations; note that a political decision is needed to be able to use even anonymised corpora.

Obviously there is a need for processing the variety of languages used on social media sites, but the resources are scarce. Estonians do not use Twitter much; Facebook is more popular, but using Facebook data is problematic.

[6] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

As for resources for sentiment analysis, within the Horizon-2020 project EMBEDDIA,[7] a dataset for hate speech detection, containing ca. 3 million annotated comments, was created and the Institute for Estonian Language has some resources for emotion detection.[8]

Estonian is included in the multilingual resources of the EU languages.

As for specific bi- and multilingual data, the Estonian-Russian language pair is of particular interest for Estonia, as 29% of the population in Estonia speak Russian as their first language.

Finland and Latvia are our geographical neighbours, so Estonian-Latvian and Estonian-Finnish machine translation is needed for practical purposes also. There is little parallel Estonian-Finnish or Estonian-Latvian data that is a result of direct translation between these language pairs; most of the texts have been created by translating an English original into Estonian, Latvian or Finnish. There is also too little test data for machine translation with Estonian as the source language.

We have at least a minimal necessary amount of audio resources for Estonian (17 corpora containing both text and audio data), but more and/or bigger special corpora are needed: children's and senior's speech, accented speech, and also speech of people having specific medical conditions (Parkinson's disease, Alzheimer's disease, dementia).

We also need more audio data for natural and noisy communication situations: spontaneous conversations, spontaneous meetings etc.

At the moment, there is no corpus of Estonian sign language, but the need for developing resources for sign language has been recognised by the government.[9]

Lexical-conceptual resources of Estonian are mostly lexicons, dictionaries and machine-readable dictionaries, as well as terminological databases. The majority of them are compiled and continuously upgraded by the Institute of Estonian Language, which has recently started to consolidate its lexical resources into a language portal *Sõnaveeb*[10] (Word Web) (Koppel et al., 2019), with the information displayed there coming from a Dictionary Writing System *Ekilex*.[11] As of February 2021, Ekilex contains about 80 lexical databases: general as well as specialised dictionaries.

An important lexical-conceptual resource is Estonian Wordnet.[12] As of October 2021, the Estonian Wordnet contains about 91,700 concepts (synsets) and continues to grow.

We lack a Framenet-type lexical resource for integrating syntax and semantics and for describing verb valency in Estonian.

There is only one type of full-coverage rule-based computational grammar for Estonian: Constraint Grammar,[13] which contains rule-sets and lexicons for morphological disambiguation, clause segmentation, syntactic function labeling and dependency structure. In both Grammatical Framework[14] and Giellalt,[15] there is a rule- and lexicon-based morphological model, with the lexicon based on that of Vabamorf.[16]

As for massive monolingual models, there is EstBERT,[17] a pretrained BERTBase model exclusively trained on an Estonian cased corpus. In addition, Estonian RoBERTa[18] and ELMo[19] models are exclusively trained on Estonian data. The newest monolingual model is a large-

[7] http://embeddia.eu
[8] http://peeter.eki.ee:5000/?lg=en
[9] https://www.hm.ee/sites/default/files/htm_eesti_keele_arengukava_2020_a4_web_en.pdf
[10] https://sonaveeb.ee
[11] https://ekilex.eki.ee
[12] https://www.cl.ut.ee/ressursid/teksaurus/
[13] https://github.com/EstSyntax/EstCG
[14] https://www.grammaticalframework.org
[15] https://giellatekno.uit.no
[16] https://github.com/Filosoft/vabamorf
[17] https://huggingface.co/tartuNLP/EstBERT
[18] https://huggingface.co/EMBEDDIA/est-roberta
[19] https://www.clarin.si/repository/xmlui/handle/11356/1277

size GPT2 model,[20] trained from scratch on 2.2 billion words.

Multilingual models include XLM-RoBERTa[21] and FinEstBert,[22] among others. As for domain-specific models, there is the WIKIBert-et[23] model trained on Estonian Wikipedia.

## 4.2 Language Technologies and Tools

The existing tools cover the basics of text analysis – sentence segmentation, tokenisation, morphological analysis, syntactic parsing – for standard written language. As soon as the text deviates from the standard, the quality of the analysis decreases.

The Estonian language has a rich morphological system, so converting a word-form to its lemma using a simple stemmer is often not possible; instead, proper lemmatisation is needed. Accordingly, the basic tool for analyzing Estonian text is a morphological analyzer.

There are two morphological analyzers for Estonian, both of them also perform morphological wordform generation and have a separate disambiguation module: Vabamorf[24] and EKI morphological analyzer.[25] Both of them are open source. The rule-based Constraint Grammar EstCG also includes a module for morphological disambiguation.

For syntactic analysis, there are the rule-based Constraint Grammar[26] surface syntax and dependency syntax modules and dependency parsing models trained on the Estonian UD treebanks (Stanza,[27] SpaCy,[28] UDPipe[29]).

The EstNLTK Python library[30] (Laur et al., 2020) contains open source tools for Estonian NLP. The corresponding pipeline starts from tokenisation and ends with syntactic analysis and information extraction (named entity recognition, grammar-based address recognition etc).

TEXTA Toolkit[31] is a program that provides resources for text analytics or solutions based on the latter; the content of the toolkit can be configured according to the needs of the customer.

Although the quality of speech processing tools and services is far from the quality of those for English, the situation for Estonian is quite good, at least for “ordinary” speech, i. e., while the speaker is speaking Estonian as their first language and has no specific health conditions and there is little background noise. For speech recognition there is, for example, TalTech’s speech recognition system[32] with available source code and also more specific services, e. g., subtitles for Estonian live broadcasts using ASR[33] or a rich transcription system for the Estonian Parliament.

There are also several models for speech synthesis,[34] including a neural network-based one.[35]

The EU’s translation tool eTranslation provides machine translation services for Estonian. Estonian is a featured language in Google Translate; Microsoft Translator provides a text and

[20] https://huggingface.co/tartuNLP/gpt-4-est-large
[21] https://huggingface.co/docs/transformers/model_doc/xlmroberta
[22] https://huggingface.co/EMBEDDIA/finest-bert
[23] https://huggingface.co/TurkuNLP/wikibert-base-et-cased
[24] https://github.com/Filosoft/vabamorf
[25] http://www.eki.ee/tarkvara/analyys/
[26] https://github.com/EstSyntax/EstCG
[27] https://stanfordnlp.github.io/stanza
[28] https://github.com/EstSyntax/EstSpaCy
[29] https://lindat.mff.cuni.cz/services/udpipe/
[30] https://github.com/estnltk/estnltk
[31] https://github.com/texta-tk/texta
[32] https://tekstiks.ee
[33] https://github.com/alumae/kiirkirjutaja
[34] http://www.eki.ee/heli/index.php
[35] https://neurokone.ee

speech translation service for Estonian as well as an offline translation pack.

The translation quality depends on the domain: general domain texts are translated better, translating a text belonging to a specialised domain may give worse results. In general, these translation services are sufficient for getting a general understanding about the content of a general-domain text.

However, independent MT services are important for government sector and translation agencies as they can not share their data with companies like Google.

There have been several projects to collect data and develop machine translation engines that support translating to and from Estonian. However, machine translation technology is not yet widely adopted. To tackle this, the government has initiated the central translation platform project (Tõlkevärav) – a national platform to help public and private sector companies manage their translation jobs, translation memories, and use machine translation. The analysis[36] for the project was done in 2021 and the development will follow in upcoming years. Additionally, Estonian will participate in NLTP (National Language Technology Platform) – a CEF project to create an NLP platform that includes machine translation functionality.[37]

In the field of Information Extraction and Information Retrieval, there are several NER models, as a part of EstNLTK[38] or on top of BERT[39] and also resources for time expression extraction,[40] but not for event extraction and event classification, although there have been some student projects on the subject.

The Texta toolkit[41] for terminology extraction and text analytics enables document classification, terminology extraction and topic detection.

There is little work done on language generation and summarisation for Estonian. However, the need for such tools has been recognised and first experiments have been performed.

Conversational agents or chatbots are widely used on the webpages of companies and government institutions to provide help for most common problems, and to guide clients to employees able to solve their problems. Rule-based customer support chatbots are mostly developed by private companies (MindTitan, AlphaBlues, etc) and are mostly used by large-scale private companies to alleviate their customer support workloads.

On the other hand, existing virtual assistant solutions (Alexa, Siri, etc) provide little value for Estonian as they don't understand the language nor are they not integrated with Estonian services.

In 2020 the Estonian government came forward with a vision of how digital public services should work in the age of artificial intelligence and launched an initiative called Bürokratt[42] – an interoperable network of AI applications, which enable citizens to use public services with virtual assistants through voice-based interaction.

## 4.3 Projects, Initiatives, Stakeholders

The need for LT support has been acknowledged by Estonian government agencies and policy-makers. Since 2006 there has been a series of National Programmes for Language Technology, with the current one in force until the year 2027.[43]

[36] https://wiki.rik.ee/pages/viewpage.action?pageId=82480426
[37] https://www.european-language-grid.eu/wp-content/uploads/2021/11/Project-Profile_NLTP.pdf
[38] https://github.com/estnltk/estnltk
[39] https://github.com/TartuNLP/bert-ner-service
[40] https://github.com/soras/Ajavt
[41] https://github.com/texta-tk/texta
[42] https://en.kratid.ee/buerokratt-v2
[43] https://www.hm.ee/en/activities/research-and-development/research-programmes

A report on the Estonian AI taskforce[44] was published in 2019, as well as Estonia's national AI strategy for the years 2019-2021.[45]

The Estonian Language Development Plan[46] sets out Estonia's language policy goals and development directions for the years 2021–2035. Development of Language Technology is stated as a priority. Several of the main planned activities include LT activities.

The national research infrastructure relating to LT in Estonia is the Center of Estonian Language Resources CELR[47] and the Competence Center for Natural Language Processing at the Institute of the Estonian Language[48].

Key stakeholders also include the Ministry of Education and Research, and the Ministry of Economic Affairs which are responsible for integrating LT into state information systems and also for developing an interoperable network of AI applications called *bürokratt*,[49] which enable citizens to use public services with virtual assistants through voice-based interaction. In addition, the Ministry of Justice is an active user of LT technologies.

Estonia is a member of CLARIN (Common Language Resources and Technology Infrastructure), ELRC (European Language Resource Coordination), and ELG (European Language Grid).

# 5 Cross-Language Comparison

The LT field[50] 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 never seen before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

## 5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[51] 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)

[44] https://f98cc689-5814-47ec-86b3-db505a7c3978.filesusr.com/ugd/7df26f_486454c9f32340b28206e140350159cf.pdf
[45] https://f98cc689-5814-47ec-86b3-db505a7c3978.filesusr.com/ugd/7df26f_27a618cb80a648c38be427194affa2f3.pdf
[46] https://www.hm.ee/sites/default/files/eesti_keele_arengukava_2035.pdf
[47] https://www.keeleressursid.ee/en/
[48] https://portaal.eki.ee/tegevusvaldkonnad/106.html
[49] https://en.kratid.ee/burokratt
[50] This section has been provided by the editors.
[51] 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.

  - Image/video processing (e. g., facial expression recognition)
  - Human-computer interaction (e. g., tools for conversational systems)

- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

## 5.2 Levels of Technology Support

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

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in $\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[52]

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 of 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[53] 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.

[52] 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.

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

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[54]

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.

With that being 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,[55] 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

[54] 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.

[55] 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.

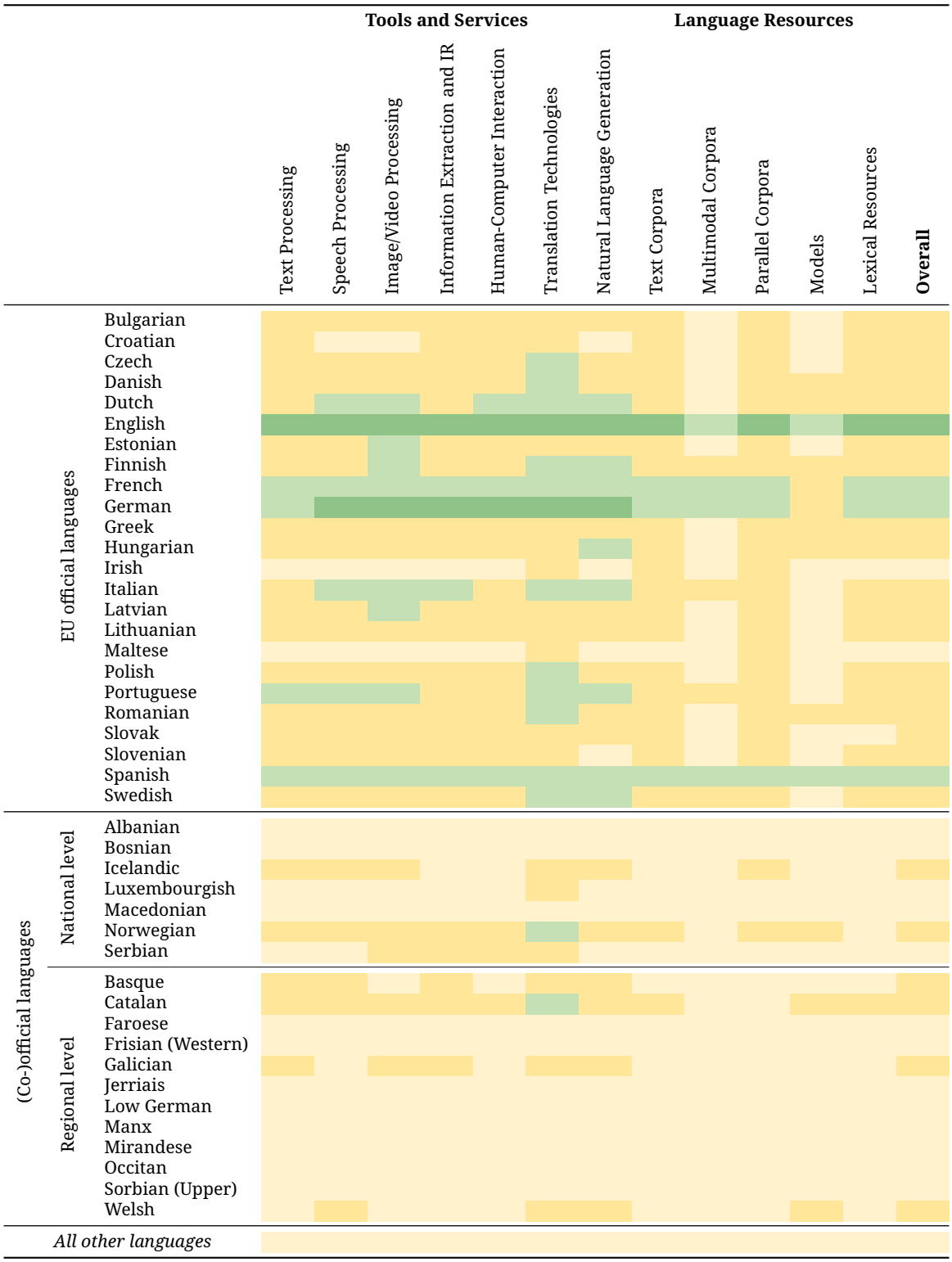


Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

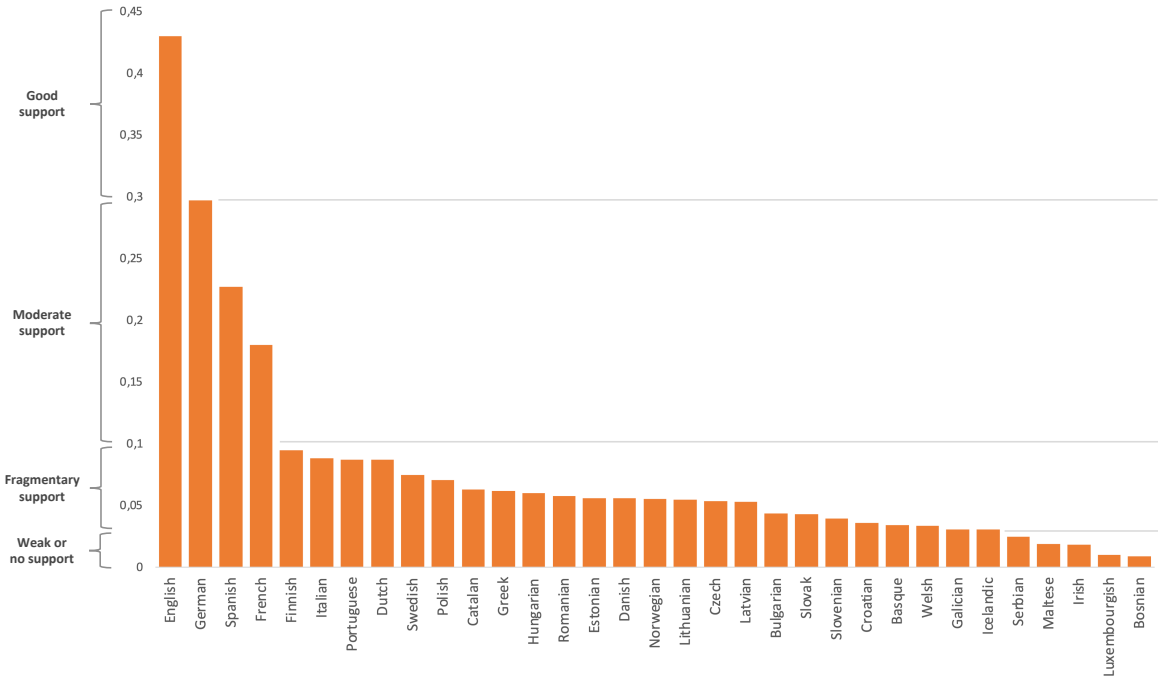


Figure 1: Overall state of technology support for selected European languages (2022)

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 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

Generally speaking, the current situation of Estonian is acceptable for a small language, but far from perfect and if measured against the LT support for English language as a benchmark, it lags severely behind.

Estonia is a small country with around one million speakers of Estonian altogether, which means that the market for language technology products for Estonian is also a small one. As might be expected, the main force driving the development of Estonian language technology has been the public sector and this has resulted in the positive fact that the resources and tools developed by state-funded projects are open source. The less desirable effect of state-funded projects is that they are mostly research and development projects, and the deliverables of those projects tend to be prototypes, not finished products.

Still, during the last decade the situation has slowly but surely been improving, as now the private sector has also engaged in creating tools and solutions for Estonian language technology.

During the last decade some fields, especially machine translation and speech technology, have advanced significantly and we have better and bigger corpora of contemporary written language and bigger treebanks, but several gaps that were identified by the Meta-Net White Paper in 2012 (Liin et al., 2012) are still there: text generation is still under-developed and we lack annotated semantic resources and tools for semantics.

The existing tools cover the basics of text analysis – sentence segmentation, tokenisation, morphological analysis, syntactic parsing – for standard written language. As soon as the text deviates from the standard, the quality of the analysis decreases significantly.

The overall quality of machine translation and especially of speech technologies are also quite satisfactory, but again only for standard language.

While talking about gaps, it is usually the case that we lack both annotated data and tools for certain tasks and, as annotating data is a time- and workforce-consuming process, it can be seen as an even bigger obstacle.

There are annotated resources available for developing the basic tools for segmentation, morphology and syntax, but again, they represent the standard written language. Accordingly, we seriously lack annotated corpora for non-standard language varieties.

There are large web-crawled corpora for Estonian, but less domain-specific corpora or, if such resources exist, they are not publicly available. Accordingly, resources need to be made available – as has been done successfully in other countries, e. g., Ireland – to persuade Estonian data-holders of the benefits of sharing such data sets.

Most of the work in Estonian LT is done at academic institutions and is project-based. Once the project is over, the developed resources are not updated any more. Accordingly, there is a need for an infrastructure for keeping these models and tools up-to-date once the project has ended so that Estonia can continuously benefit from that important work.

The general attitude in Estonia towards digitisation is definitely positive and people as well as government sector are ready and even eager to use AI and LT tools in their everyday life. On the other hand, the relatively small number of speakers (=buyers of LT products) hinders the development. Government has understood the necessity and value of Language Technology, but during the last few years, the Government's main concern has of course been the pandemic, which has overshadowed everything else. So, if Estonian citizens, the public and private sector are given access to better LT tools, they will definitely make a good use of them.

# References

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

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

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

Mati Erelt. *Estonian Language*. Linguistica Uralica Supplementary Series. Estonian Academy Publishers, Tallinn, 2003.

Kristina Koppel, Arvi Tavast, Margit Langemets, and Jelena Kallas. Aggregating Dictionaries into the Language Portal Sõnaveeb: Issues With and Without Solutions. 2019. doi: 10.5281/zenodo.3612931. URL https://doi.org/10.5281/zenodo.3612931.

Sven Laur, Siim Orasmaa, Dage Särg, and Paul Tammo. EstNLTK 1.6: Remastered Estonian NLP pipeline. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7152–7160, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.884.

Krista Liin, Kadri Muischnek, Kaili Müürisep, and Kadri Vider. *Eesti keel digiajastul – The Estonian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30784-3. Available online at http://www.meta-net.eu/whitepapers.

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.

David Smahel, Hana Machackova, Giovanna Mascheroni, Lenka Dedkova, Elisabeth Staksrud, Kjartan Ólafsson, Sonia Livingstone, and Uwe Hasebrink. Eu kids online 2020: Survey results from 19 countries. 2020. URL https://doi.org/10.21953/lse.47fdeqj01ofo.

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.