EUROPEAN LANGUAGE EQUALITY

D1.26

Report on the Norwegian Language

Authors	Kristine Eide, Andre Kåsen, Ingerid Løyning Dale
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	D1.26
Deliverable title	Report on the Norwegian Language
Type	Report
Number of pages	30
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Kristine Eide, Andre Kåsen, Ingerid Løyning Dale
Reviewers	Maria Giagkou, Annika Grützner-Zahn
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

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
AMR	Abstract Meaning Representation
API	Application Programming Interface
CG	Constraint Grammar
CLARINO	Common Language Resources and Technology Infrastructure Norway
DLE	Digital Language Equality
ELE	European Language Equality *(this project)*
ELRC	European Language Resource Coordination
HPC	High-Performance Computing
LBK	Leksikografisk bokmålskorpus
LR	Language Resources/Resources
LT	Language Technology/Technologies
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
NCC	Norwegian Colossal Corpus
NDC	Nordic Dialect Corpus
NDT	Norwegian Dependency Treebank
NLB	Norsk Lyd- og Blindeskriftsbibliotek (Norwegian Library of Talking Books and Braille)
NLP	Natural Language Processing
NLN	The National Library of Norway
NoReC	Norwegian Review Corpus
NorNE	Norwegian Named Entities
NoWaC	Norwegian Web as Corpus
NPSC	Norwegian Parliament Speech Corpus
NST	Nordic Language Technology Holding AS
NTNU	Norges Teknisk-Naturvitenskapelige Universitet
OBT	Oslo-Bergen Tagger
OCR	Optical Character Recognition
OMC	Oslo Multilingual Corpus
SR	Speaker Recognition
UD	Universal Dependencies
UiO	University of Oslo

Abstract

Language technology tools and services have greatly increased in Norway in recent years, as have the linguistic resources that are needed to make them work. In the past 10 years, we have adopted new or improved versions of machine translation, speech technology, chatbots and digital assistants, and machine learning has improved. Nevertheless, language technology for both written standards of the Norwegian language – the majority Bokmål and minority Nynorsk – is nowhere near the same level as that of major European languages such as English, German, French and Spanish.

One of the purposes of this report is to identify what is needed for Norwegian language technology to reach the level of major languages. Which basic resources should we invest in? Which tools are either completely lacking or in need of improvement? And what kind of research and development is required for both standards of Norwegian to remain viable in the future, also within fields such as technology, online services and higher education?

Norwegian lacks the big data that machine learning requires: this applies to both standards, but especially Nynorsk. Norwegian speech recognition understands standardised Eastern Norwegian best, and dialect recognition is far from good enough. We lack domain-specific language data that enables language technology to work within specific domains. It is crucial that the work to provide basic resources continues and to have continued funding for this type of infrastructure. Existing data must be collected, new data sets must be produced, and everything must be made available for further use in language technology.

Awareness of the differences between Nynorsk and Bokmål is low among operators outside Norway's borders. Even though Norwegian is found in large, multilingual resource collections that are used to create language models, and even though Norwegian is available as a language choice also on large international platforms, it is first and foremost Bokmål tools and services that are developed.

An important point here is to raise the general awareness of what language technology is, what it can be used for, what problems it can solve and what resources can be reused to create good language technology tools for Norwegian. One area where we see a steady increase from 2012 to today is resources used for automatic translation. This is a result of a European investment in machine translation under the CEF digital programme , where one of the tasks was to inform the public sector about the value of data in the form of translation memories and the subsequent collection of such memories from public enterprises.

As of today, there is no research programme in Norway aimed specifically at language technology. Nevertheless, several Norwegian projects are in the process of filling some of the gaps that have been identified. There are projects on improving speech recognition, sentiment analysis, anonymisation, text collection and methods for training systems with smaller amounts of data. In order for Norwegian language technology to have the best possible conditions, participation in international projects and initiatives should be facilitated in order to utilise the transfer value between languages, while at the same time there should be an incentive to research and develop parallel tools for Bokmål and Nynorsk. Although Bokmål and Nynorsk are quite similar and there is a transfer potential between them, it is necessary to create parallel editions for the languages. This must be taken into account in the financing of Norwegian language technology and in the purchase of language technology solutions. The public sector in Norway should use its purchasing power to ensure parallel Bokmål and Nynorsk versions when purchasing digital solutions from both small suppliers and large, international companies.

Norsk samandrag

Språkteknologi er i rivande utvikling. I Noreg har vi i dei siste åra vore vitne til ein kraftig auke både i mangfaldet av språkteknologiske verktøy og i bruken av dei. Denne teknologiutviklinga har konsekvensar for språket. Prop. 108 L Lov om språk (Ministry of Culture, 2020) slår fast at norsk må bli brukt i digitale tenester og produkt om norsk skal vere eit samfunnsberande språk i åra som kjem.

Denne rapporten kartlegg grunnlagsressursar for språkteknologi og verktøy som er utvikla for norsk. Rapporten er ein av 32 rapportar om språkteknologi for ulike europeiske språk. Til saman gjer desse rapportane det mogleg å jamføre norsk språkteknologi med språkteknologi for dei andre europeiske språka. I jamføringa ligg norsk omtrent på same plass som då den førre, tilsvarande rapporten kom ut for ti år sidan, til tross for at både grunnlagsressursane og verktøya som er tekne i bruk, har vorte fleire og betre.

Sidan 2012 er det teke i bruk nye eller forbetra utgåver av maskinomsetjing, taleteknologi, praterobotar (chatbots) og digitale assistentar. Utviklinga innanfor maskinlæring har bidrege til det. Likevel er språkteknologi for norsk langt frå det nivået som dei store europeiske språka som engelsk, tysk, fransk og spansk ligg på. Også språkteknologi for desse språka har hatt ei rivande utvikling dei siste åra. Derfor kjem norsk språkteknologi omtrent midt på treet når ein jamfører med desse språka og andre europeiske språk.

Eit av føremåla med denne rapporten er å identifisere kva som skal til for at norsk språkteknologi skal kome opp på nivå med dei store språka: Kva grunnlagsressursar bør vi satse på, kva verktøy manglar eller må bli betre, kva trengst av forsking og utvikling for at norsk skal vere eit samfunnsberande språk òg i framtida?

Språkteknologi trengst for at digitaliseringa av norsk offentleg og privat sektor skal fungere. Språkteknologi blir brukt til tekstanalyse, taleattkjenning og tekst-til-tale-system, automatisk omsetjing, nettsøk, automatisk referatskriving, tekstsamandrag og i praterobotar og digitale assistentar. Han er ein føresetnad for oppfylling av krava til universell utforming. Språkteknologi er dessutan ein viktig komponent i kunstig intelligens.

Norsk (både bokmål og nynorsk, men særleg nynorsk) manglar dei store datamengdene som maskinlæring krev. Norsk taleattkjenning forstår standard austnorsk best, og dialektattkjenninga er langt frå god nok. Vi manglar domenespesifikke språkdata som gjer at språkteknologi kan fungere innanfor einskilde fagområde. Eksisterande data må samlast inn, nye datasett må produserast, og alt må gjerast tilgjengeleg for vidare bruk i språkteknologi. Det er svært viktig å halde fram med det arbeidet som blir gjort i Språkbanken med å skaffe grunnlagsressursar. Også annan språkteknologisk infrastruktur som finst i norske forskingsinstitusjonar, må førast vidare.

Operatørar utanfor Noreg har lite kunnskap om den norske språksituasjonen med to jamstilte skriftspråk. Sjølv om norsk finst i store, fleirspråklege ressurssamlingar som blir brukte til å lage språkmodellar, og sjølv om norsk er tilgjengeleg som språkval òg på store internasjonale plattformer, er det først og fremst bokmålsressursar som blir utnytta. Til dømes vil ein språkmodell som blir opplært på eit stort norsk korpus som inneheld tekst på begge skriftspråk, primært vere ein modell for bokmål, sidan bokmål utgjer det prosentvis største tekstgrunnlaget.

Per i dag finst det ikkje noko forskingsprogram som er retta spesielt mot språkteknologi. Likevel er fleire norske prosjekt i ferd med å bøte på nokre av dei manglane vi har peika på i denne rapporten. Det er mellom anna prosjekt som arbeider med betre taleattkjenning, sentimentanalyse, anonymisering, tekstinnsamling og metodar for å lære opp system med mindre datamengder. Skal norsk språkteknologi få best moglege føresetnader, bør det leggjast til rette for at norske aktørar deltek i større internasjonale prosjekt. Eit område der norsk har kome på eit høgare nivå jamført med nivået i rapporten frå 2012, er ressursar og verktøy som blir brukte til automatisk omsetjing. Dette er eit resultat av ei europeisk satsing

på maskinomsetjing under CEF-digitalprogrammet. Det bør òg finnast insentiv til å forske på og utvikle parallelle verktøy for bokmål og nynorsk. Forskingsrådet som statleg organ må ved tildelingar ta omsyn til formuleringa i Prop. 108 L Lov om språk (Ministry of Culture, 2020) om at det har «eit særleg ansvar for å fremje nynorsk som det minst bruka norske skriftspråket.»

Det må lagast parallelle utgåver av språkteknologiske verktøy for dei to norske skriftspråka. Så lenge språkteknologien fungerer dårlegare på nynorsk enn på bokmål, vil det ikkje vere mogleg å nå hovudmålet i den norske språkpolitikken om at desse språka skal vere jamstilte. Det må takast omsyn til dette når norsk språkteknologi skal finansierast, og når språkteknologiske løysingar skal kjøpast inn. Det offentlege Noreg bør bruke innkjøpsmakta si til å sikre parallelle bokmåls- og nynorskversjonar når dei kjøper digitale løysingar både frå små leverandørar og frå store internasjonale selskap.

Ei oppsummering av dei tiltaka som trengst for at norsk språk skal vere samfunnsberande på digitale flater også i framtida:

- Det må lagast verktøy og skaffast ressursar som manglar i dag, inkludert større mengder tekstdata for nynorsk, fleire domenespesifikke data, leksikalske og terminologiske ressursar (spesielt for nynorsk), og dessutan taledata som dekkjer dialektar og nynorsk, og verktøy for semantisk analyse.
- Slike ressursar og verktøy bør gjerast tilgjengelege under så opne lisensar som mogleg, slik at ein kan sikre ombruk av dei.
- Det må lagast datasett som tillèt tekstanalyse over setningsnivå, som til dømes eit koreferanse-korpus.
- Medvitet om verdien av språkdata må aukast.
- Norske institusjonar må halde fram med å delta i internasjonale forskingsprosjekt og andre prosjekt som fokuserer på språkteknologi, som ELE, ELG og ELRC.
- Det må setjast av nok midlar til utvikling av språkspesifikk språkteknologi for norsk.
- Offentleg sektor må ta det ansvaret som språklova legg på sektoren, og sikre parallelle versjonar av norsk språkteknologi i offentlege innkjøp. Offentleg sektor må lage standardformuleringar som skal brukast ved offentlege innkjøp, slik at sektoren sikrar seg retten til språkressursar som kjem frå omsetjingar og andre tenester.
- Kvaliteten på norsk språkteknologi bør kartleggjast for at ein skal kunne jamføre nynorsk og bokmål og kunne vurdere dialektforståinga i taleattkjenning.

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 identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

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

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

2.1 General Facts

Norwegian is a North Germanic language, spoken by approximately five million people in Norway.[3] All children, except some speakers of indigenous and minority languages, learn Norwegian as their first language at school.

There is great dialectal variation in Norway, and people tend to speak their own dialect. Unlike other official European languages, there is no official standard for spoken Norwegian. The pronunciation of the written languages in the Norwegian Broadcasting Corporation (NRK) is often regarded as a standard, albeit an unofficial one. Dialects have a much higher prestige than in the other Scandinavian countries, and Norwegians generally expect their dialect to be understood by other Norwegians. During the last 50 years, there has been a steady increase in the use of dialects in an expanding variety of contexts. That said, the most commonly spoken variety is often referred to as Standard Eastern Norwegian.

While there is no official standard for the spoken language, there are two official standard written Norwegian languages, Bokmål and Nynorsk. With some exceptions, children are taught both varieties at school, and the schools can choose one of them as the main language, teaching the other variety as the "side language". Statistics from the primary schools show that 11.2% of the pupils learn Nynorsk as their main language. However, many children who are taught Nynorsk in school change from Nynorsk to Bokmål when they start high school. The percentage of pupils who learn Nynorsk as their main language has decreased gradually since 1940, so we might expect a larger proportion of Nynorsk users among older age groups. It may be reasonably realistic to assume that about 500,000 Norwegians use Nynorsk as their first written language.

The linguistic differences between Bokmål and Nynorsk, in regard to vocabularies, spelling, morphology and syntax, are rather small. Nevertheless, for most types of language technology, such as machine translation, chatbots, spelling checks, speech-to-text and text-to-speech, separate tools are needed for each language.

One peculiarity to Norwegian is the large formal variation in both written languages. Both are standardised to reflect some dialectal variation, such as freedom in the gender of some nouns and spelling variations. In combination with highly productive compounding, one single word can reach a relatively high number of different spellings. This variation is a challenge for language technology.

In relation to speech technology, the dialectal variation, which is high compared to the variation in the written languages, is a challenge. Most Norwegian dialects have contrastive

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

[2] https://european-language-equality.eu

[3] With some additional speakers in the Norwegian diaspora in the US and South America. There has been no census on the number of native speakers.

pitch, often called toneme 1 and 2, that constitute minimal pairs. The tonemes and the homographic spelling of some words that can only be distinguished by pitch has been another challenge for text-to-speech systems.

A new language act

The new Norwegian language act, effective from 1 January 2022, recognises Norwegian as the official language in Norway,[4] and gives Bokmål and Nynorsk equal status as separate written languages. Previously, they were regarded as two varieties of Norwegian. The new act continues previous regulations of the use of Bokmål and Nynorsk in the public sector and affirms that correspondence between citizens and the administration at state level shall be in the language of the citizen's choice and that a minimum of 25% of publications from state level bodies shall be in one of the languages. Few state institutions reach the 25% target in Nynorsk today, and one purpose of the law is to ensure that public bodies take responsibility for the use, development and strengthening of both Bokmål and Nynorsk. This includes a special responsibility to strengthen Nynorsk as the least-used written Norwegian language.

2.2 Norwegian in the Digital Sphere

Norway is a highly digitalised society. By 2021, 99% of the population between 16 and 79 had used the internet in the previous three months,[5] and there are more than 830,000 domain names under .no.[6]

The Norwegian public sector communicates with its citizens on digital platforms. Because language technology is a prerequisite for digitalisation to work, there is a growing awareness that well functioning Norwegian language technology is fundamental to the democratic rights of the citizens.

The latest public reports on language (Ministry of Culture, 2020), digitalisation (Ministry of Local Government and Regional Development, 2019) and artificial intelligence (Ministry of Local Government and Regional Development, 2020) all mention language technology as a prerequisite for digitalisation and digital communication, and for Norway to be able to capitalise on artificial intelligence. There is also a requirement that language technology used in the public sector must support Norwegian. Because a Norwegian citizen can choose their preferred language for communication with the public sector on the national level, a chatbot, for instance, must be able to understand questions and answer them in both Nynorsk and Bokmål. Digital forms must be available in both languages simultaneously.

The digitalisation of the public sector has thus led to an increase in the use of Norwegian language technology as well as to increased awareness of what language technology is and what it can be used for. Unfortunately, the technologies still do not work as well in Norwegian as they do in English, and the quality is lower for Nynorsk than it is for Bokmål. Internationally, Norwegian is an option on large platforms such as Google and Facebook and in Apple and Microsoft products. The Norwegian market, with its five million speakers, is not always considered large enough for the localisation of new products. When a Norwegian version exists, few international developers have parallel Nynorsk and Bokmål versions.

Traditionally, language issues sit under the Ministry of Culture. Because the use of plain language and automated communication such as chatbots make interaction with the public more efficient, the economic aspects of language and language technology have caught

[4] Sami languages, Norwegian minority languages and Norwegian sign language are also recognised in the new act. These languages do not fall within the scope of this report.

[5] SSB, Jan. 2022: https://www.ssb.no/en/teknologi-og-innovasjon/informasjons-og-kommunikasjonsteknologi-ikt/statistikk/bruk-av-ikt-i-husholdningene

[6] NORID, Dec. 2021: https://www.norid.no/en/om-domenenavn/nokkeltall/

the attention of other sectors within public administration. The same is true for machine translation, which has received attention through the CEF programme and the eTranslation building block. The digitalisation of the public sector and the need for language technology, as well as the success of plain language initiatives, has led to a growing interest in language in general, and language technology in particular, in other sectors.

While the Norwegian language is not threatened by digital extinction, certain domains are dominated by English, and in some areas there is a lack of proper Norwegian terminology. According to the status report on the Norwegian language (Language Council of Norway, 2021), certain domains, in particular technical ones, are at risk of being taken over by English. One purpose of the new language act is to ensure that each sector of the public administration takes responsibility for the Norwegian language within their domain. The requirement to "use, develop and strengthen" both languages means that public bodies must ensure the development of terminology within their field of expertise.

All three strategies on language, digitalisation and artificial intelligence mentioned above describe the link between language technology (LT) and language resources (LRs) for the development of Norwegian language technology. They also mention the most important investment in basic language technology resources, the Language Bank at the National Library (see Section 4).

Digitalisation of the public sector has increased the amount of data that can be used for language technologies. Some attention has been given to gathering these public data, such as translations, terminology and textual resources from public administrations and depositing them in the Language Bank. The increase in data availability from 2018 to 2021 has been substantial. Even so, awareness of what language data is and what it can be used for needs to be raised in both the public and private sectors.

3 What is Language Technology?

Natural language is the most common and versatile way for humans to convey information.[7] We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

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

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

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the

[7] 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).

1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning (ML), rule-based systems have been displaced by data-based ones, i. e. systems that learn implicitly from examples.

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

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

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e. the generation of speech given a piece of text, Automatic Speech Recognition, i. e. the conversion of speech signal into text, and Speaker Recognition (SR).
- **Machine Translation**, i. e. the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.
- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e. the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction** which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). A very popular application within this area are conversational agents (better known as chatbots).

LT is already fused in our everyday lives. As individual users we may be using it without even realizing it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just very few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It

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

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[8]

4 Language Technology for Norwegian

The report The Norwegian Language in the Digital Age (Smedt et al., 2012) concluded that Norwegian had five urgent needs: improved licensing conditions and standardisation of existing basic tools and resources; creation of missing basic tools and resources; research on automated linguistic analysis, as well as integrating statistical and rule-based LT; higher visibility of research results; and long-term funding strategies for the development of resources for both written standards and minority languages

To respond to these needs, the Language Bank, *(Språkbanken)*[9] was established in 2010 with the aim of making language technology resources available to both the public and private developers, for commercial use as well as for research (Ministry of Culture, 2020). The Language Bank was reviewed in 2018 and was granted extra funding from the state budget on a yearly basis, and since then the resources available to Norwegian language technology have substantially increased. The Language Bank is a key tool in Norwegian language policy, and the increase in publicly available tools, lexical resources, and text and speech corpora is mostly due to the political will to fund their development and maintenance, as well as greater popular demand for higher-quality language technology services.

The importance of open data has become a mantra for AI development. Because of the need for language technology in the digitalisation of the public sector, efforts have been made to ensure that all public language data that can be made openly available are indeed shared through the Language Bank. Due to the low awareness of what language data is, the National Library, the Norwegian Language Council (*Språkrådet*) and the Norwegian Digitalisation Agency (*Digitaliseringsdirektoratet*) have provided joint guidelines for identifying and sharing language data for further use.[10]

Besides the lack of awareness and identification, the two main hurdles for sharing language data are copyright issues and the implementation of the GDPR. For instance, valuable large speech corpora, such as subtitled television shows and transcribed radio programmes from public media, are not publicly available for further use, because of copyright restrictions. Journalism is exempt from certain aspects of the privacy laws, but when we turn journalistic material and news articles into linguistic resources, they are no longer exempt and are consequently subject to privacy restrictions. There is an ongoing process of finding an acceptable way of sharing such data without violating privacy concerns, which involves content owners, lawyers and the LT community. The Norwegian Digitalisation Agency has

[8] 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).

[9] https://www.nb.no/sprakbanken/en/sprakbanken/

[10] https://www.digdir.no/datadeling/sprakdata-korleis-kan-vi-hauste-og-dele/2367

established the Norwegian Resource Centre for Sharing of Data. Even though it does not specialise in language data, it can help with the legal framework and interpretation of it, which is one of the obstacles for data sharing in general.

4.1 Language Data and Tools for Norwegian Bokmål

The overall accessibility of language resources for Bokmål is fairly good. Size and contemporaneity are in place for unstructured and semi-structured data. With what is available, and with good linguistic insight, one can build several specialised applications and services. However, the domain coverage is limited since this has only become an issue very recently and some types of tools and resources either need updating or are non-existent altogether.

Monolingual Text Corpora

The National Library of Norway (NLN) has digitalised most of its collection. This includes books, magazines, music, film, radio and TV programmes, pictures, photographs, theatre material, maps, posters, and newspapers. Some documents date as far back as the 12th century. This digital literary archive is available to the public online in a view-only format. Optical character recognised (OCR) texts that are no longer copyright-restricted are available via a corpus-building API provided by the library. Thus, for the purposes of quantitative analysis of Norwegian literature, it is possible to build one's own corpus, selecting texts based on document metadata.

For the purposes of training language models and developing language technology, the Norwegian Colossal Corpus (NCC) is the largest text corpus in Norwegian. We include it here as a monolingual corpus because the vast majority (83%) is in Bokmål, although it also contains 12% Nynorsk text data and strains of other languages, like English and Danish (Kummervold et al., 2021).[11] It includes text data from the majority of the NLN's digitally archived, non-copyrighted books and newspapers, as well as online newspaper text from the period 1998-2019, government and public reports, parliament procedures, and legal documents from lovdata.no. Other large corpora like Målfrid, subtitles from OpenSubtitles, Norwegian Wikipedia (both Bokmål and Nynorsk), and the Norwegian parts of OSCAR and MC4 are also included. While the size of this corpus is tremendous with 18.4 billion tokens, the full corpus is not available under the same open access licence. Modern books are still copyrighted and therefore only available to the researchers at the National Library of Norway,[12] and some of the corpora that are included in the NCC have restricted or proprietary access. In terms of contemporaneity, the full time span 1814-2020 is covered and should therefore reflect historical changes and variation in spelling up until the present day.

Among large Bokmål corpora, some already listed as part of the NCC, we find Norwegian Web as Corpus (NoWaC), Norsk aviskorpus (Norwegian Newspaper Corpus), Habit, Målfrid, and Leksikografisk bokmålskorpus (Lexicographic Bokmål Corpus, LBK). Most of these corpora contain a good deal of web texts and to a certain extent some overlap, with the exception of the LBK, which contains excerpts from fiction only. They constitute enough data in Norwegian to train large contextualized models like BERT and T5, which we cover in Section 4.1 below.

Dyvik et al. (2016) present a treebank based on parsing with NorGram called NorGramBank. NorGram is a grammar based on Lexical Functional Grammar (LFG) and lists 380

[11] Danish was the official written language in Norway during Danish rule (1537-1814).

[12] Books that were written by an author who is still alive, or who has been dead for less than 70 years, do not have an open access licence: https://lovdata.no/lov/2018-06-15-40/§11

complex syntactic rules (Dyvik et al., 2016, p. 3555).[13] Contrary to other treebanks for Norwegian, it contains both fiction and non-fiction.

Of the smaller annotated corpora, one of the most widely used resources for Norwegian, not counting grammar correction tools and programmes (see Smedt et al. (2012) for a review of grammar correction), is the Norwegian Dependency Treebank (NDT) (Solberg et al., 2014). NDT contains manually corrected and human assessed annotations of dependency-style syntax, morphosyntax and part-of-speech and has been broadly available to practitioners of all levels libraries like HuggingFace, spaCy, Stanza and trankit. With the rapid development of machine learning and more specifically deep learning for language technology, it has been in widespread use in these Python libraries.

A named entity annotation layer has also been added to the NDT under the name Norwegian Named Entities (NorNE) and a coreference layer is in progress. These datasets contain annotations for the same text data as the NDT. The spread of the NDT is much due to the initiative Universal Dependencies, which has striven toward a common annotation standard for treebanks. The NDT contains mostly newspaper texts, and some political documents, parliament minutes and blog texts, and therefore covers several genres, but it is limited in domain-specific topics or terminology beyond political discourse. In the newspaper part, there are sports articles, essays, interviews and more.

The project Sentiment Analysis for Norwegian Text (SANT) led by the Language Technology Group (LTG) at the University of Oslo has released several datasets for sentiment analysis and negation resolution.[14] Their corpora are based on reviews, and their biggest corpus is the Norwegian Review Corpus (NoReC). While NoReC contains both Bokmål and Nynorsk text, the Bokmål part is the larger one.

NorDial (Barnes et al., 2021) is another corpus provided by the LTG. The corpus consists of tweets classified into four possible categories: Bokmål, Nynorsk, dialect or mixed. This is one of few resources that consider both user-generated data and written dialect use.

Bi- and multilingual text corpora

Norwegian is a small language, yet large enough to often be included in multilingual projects such as ParaCrawl and CommonCrawl. Norway does not have access to the same amount of parallel data from the European institutions as the EU member states. Even so, the ELRC initiative, which Norway participates in, has contributed to a growing awareness of the reusability of translations. As a result, public administrations with in-house translation services, such as the Norwegian Ministry of Foreign Affairs, the Norwegian Maritime Authority, and EFTA, have contributed significant collections of Bokmål-English parallel data. Translation memories from smaller administrations have been collected through professional translation services. There are over three million translation units from different areas of public administration in the Language Bank's resource catalogue and the ELRC-SHARE repository, which can be used to train machine translation systems.

Official websites of Norwegian organisations, e./,g., the Institute of Public Health, the Courts of Norway, Norway's Government, etc., are the sources of several aligned parallel corpora. Nevertheless, only very few translation memories exist between Bokmål and other languages than English. The Norwegian-Spanish Parallel Corpus, The RuN-Euro corpus (Norwegian to Russian) and the Oslo Multilingual Corpus (OMC) are some examples.

Recently, the PRINCIPLE project aimed to collect and develop linguistic resources for Irish, Norwegian, Croatian and Icelandic, for domain-specific machine translation in the legal domain (see section 4.4). The resources that have come out of the PRINCIPLE project will be valuable for machine translation systems from English to Bokmål, but there has also been

[13] https://clarino.uib.no/iness/lfg-grammars
[14] https://github.com/ltgoslo

an effort to collect a Bokmål-Nynorsk parallel corpus based on textbooks. This is possible due to the Norwegian Education Act, which states that all textbooks in Norway have to be available in both languages.

Multimodal corpora

There are several corpora aimed at speech technology development for Norwegian.

The Norwegian Parliament Speech Corpus (*Stortingskorpuset*, NPSC) contains 140 hours of recordings of parliamentary procedures from 2017 and 2018. The recordings are aligned with the official, manually written minutes (1.2M tokens) and covers dialectal variation. The parallel text references are in both Norwegian standards, Nynorsk and Bokmål, and they have been automatically translated and manually corrected. The corpus is made publicly available and maintained by the Language Bank.

The Nordic Dialect Corpus (NDC) contains spontaneous speech recordings, audio and video, covering dialectal variation from 438 informants from 111 places across the whole country, as documented by a searchable map, and includes both phonetic and orthographic (Bokmål) transcriptions. The orthographic transcriptions are grammatically annotated, and are searchable via the project's web page.[15]

Adding to the list of multimodal corpora with wide dialectal coverage, NB Tale is made up of three modules: speech read from a manuscript by native speakers of Norwegian, recordings of non-native speakers reading from a manuscript, and spontaneous speech from 380 speakers from 24 different dialectal regions. The corpus was developed by the private company Lingit in order to train speech recognition systems. The Tuva speech database was intended for automatic dictation systems, and contains 24 hours of speech from 40 speakers, 36 of which have a dialect closest to the Bokmål written standard, and 4 closest to Nynorsk. About 70% of the recordings are read from manuscripts.

The dialectal variation in Norway is a challenge to speech technology. Most speech technology for Norwegian is made with or for the Standard Eastern variety, which is closely linked to Norwegian Bokmål. More dialectal data is needed to ensure that all Norwegians can use speech technologies without having to conform their language to a standardised variety they do not necessarily master. One of the main goals in Norwegian language policy is to support the established linguistic diversity of the spoken language, hence citizens should not have to change their spoken language in the face of language technology.

Lexical/conceptual resources

The most important lexical resource for Norwegian is Norsk ordbank (the Norwegian Word Bank), a lexical database for Norwegian Bokmål reflecting the official standard orthography as defined in the Norwegian dictionary *Bokmålsordboka*.[16] This dictionary is jointly owned by the Norwegian Language Council and the University of Bergen, where it is maintained and hosted. Both resources are freely available for download and use in language technology.

Many of the larger world languages have a WordNet. WordNet is a lexicon that relates words to concepts, and the Norwegian WordNet covers both Bokmål and Nynorsk.

The national terminology portal *Termportalen* aims to gather Norwegian terminology in one place.[17] At the time of writing, there is no requirement for contributors to give a free licence to the terms and termlists they provide. The site is not designed for the direct downloading of its resources, but rather as a dictionary with a search bar. An easily accessible download function would enhance its potential as infrastructure for language technol-

[15] https://tekstlab.uio.no/glossa2/ndc2
[16] https://ordbokene.no
[17] https://term.uib.no

ogy. Other resources are the SNORRE Terminology Database for technical terms, and the EU termbase from the Ministry of Foreign Affairs, both downloadable from the Language Bank.

Several pronunciation lexica have also been compiled over the years. Some of the most prominent are Onomastica, NST (Nordic Language Technology Holding AS) and LINGIT, but also the Norwegian Library of Talking Books and Braille (NLB). The NST lexicon is in Bokmål, and LINGIT is in Nynorsk.

Resources that deal with semantic role labelling, such as the English FrameNet and VerbNet, or abstract meaning representation (AMR), are as good as non-existent in either written standard of Norwegian.

Models and grammars

There are two large language models with a similar model architecture as BERT (Devlin et al., 2019) for Norwegian, NorBERT[18] and NB-BERT.[19] NorBERT has been trained on the newspaper corpus Norsk Aviskorpus, Bokmål Wikipedia, and Nynorsk Wikipedia, totalling approximately 2 billion words (Kutuzov et al., 2021). NB-BERT has been trained on the Norwegian Colossal Corpus, comprising 18.4 billion words after deduplication (Kummervold et al., 2021). These models can be fine-tuned with annotated corpora to develop task-specific tools.

NorSource is a computational grammar developed at the NTNU.[20] Another grammar available for Norwegian is NorGram.[21]

Tools and Services

Natural Language Processing (NLP) tasks, such as named entity recognition, part-of-speech tagging, tokenisation, dependency parsing and sentiment analysis, are covered for Norwegian Bokmål by several coding libraries, such as SpaCy, Stanza and Trankit. The accuracy of these tools lags behind English by a few percentage points. The aforementioned libraries are actively being developed and released with open source licences.

The Oslo-Bergen Tagger (OBT) has been utilized for almost two decades. OBT is a rule-based tagger which is based on the constraint grammar (CG) formalism (i. e. a finite state transduction technique). It was primarily developed as a part-of-speech tagger, but the CG formalism easily extends to tagging morphosyntactic features, lemmas, as well as compound analysis of complex words. In recent years, a statistical disambiguation component has been added, and the whole system was renamed OBT+Stat.

During the PRINCIPLE project the amount of parallel data for Norwegian was almost doubled in the ELRC-SHARE repository and this will hopefully contribute significantly to the improvement of machine translation for Norwegian.[22]

In speech recognition and speech synthesis, most of the systems that could be deemed usable are proprietary and not freely available. The Norwegian library for talking books and braille (NLB) has three synthetic speech voices for Bokmål, Nynorsk and English, and they are used by the library to generate audio from text books. These synthetic voices are only available via the library and the material they offer, and not as a model or API.

Other commercial voices are offered by the Norwegian company Lingit, as well as from the international companies Acapela, Nuance, ReadSpeaker and Vitec MV. Microsoft provides access to their coding library to customise and use their synthetic voices for free,[23] while Google and Apple require paid accounts in order to call their APIs. eSpeak is a freely

[18] http://wiki.nlpl.eu/Vectors/norlm/norbert
[19] https://huggingface.co/NbAiLab/nb-bert-base
[20] https://github.com/Regdili-NTNU/NorSource
[21] https://clarino.uib.no/iness/lfg-grammars
[22] https://elrc-share.eu
[23] Given that one creates an account on their platform.

accessible and downloadable speech synthesis tool,[24] but the quality in terms of naturalness and intelligibility is poor for the Norwegian voice.

Similarly for speech recognition, Omilon's digital assistant Tuva is a commercial product,[25] and licences can only be purchased by organisations. Google and Microsoft provide quite good speech recognition of the Standard Eastern dialect through their digital assistant products and their platform APIs.

Norsøk is an information retrieval tool for both Norwegian Bokmål and Nynorsk provided by Nynodata AS.[26] It returns hit results for synonyms and inflected forms of the query words in both written standards. The tool is developed to be integrated with web pages in Norwegian, in databases or information storage systems, such as digital archives, or in web search platforms.

As for language generation and abstractive summarisation, neither written standard of Norwegian is covered by any available tools or services. The Python library summa,[27] can summarise text using an extractive method, i. e., it returns a limited portion (percentage or number of words) of the original text verbatim, and works for both standards of Norwegian text.

At the time of writing, 100 municipalities in Norway are using a chatbot called Kommune-Kari on their official websites. Monthly, people engage in around 80,000 conversations with the bot, according to the developers.[28] While the chatbot's performance has been criticised for giving poor responses for 1 in 10 questions,[29] the system architecture seems to be based on reinforcement learning, or at least include human-in-the-loop input, which in theory should enable the system to improve performance over time.

4.2 Language Data and Tools for Norwegian Nynorsk

Despite the relevant policies described in Section 1, Nynorsk language resources and tools are sparse compared to Bokmål.

Monolingual text corpora

Of its 600k tokens, the Norwegian Dependency Treebank (NDT), already described in section 4.1, has 300k tokens in each written standard. This equal distribution of Nynorsk and Bokmål is not an automatic guarantee for the actual use of both parts of the treebank in the development of language technology for Norwegian. SpaCy, for example, has not used the Nynorsk part of the treebank for their NLP pipeline.

The Norwegian BERT-based models are trained with both written languages, where the Nynorsk proportion is significantly smaller. To remedy the scarcity of Nynorsk texts, the Language Bank has harvested available legal documents from municipalities where Nynorsk is the main language. The material in PDF format has been converted to text and made available as a 127-million-word corpus.[30]

Bi- and multilingual text corpora

Among the Norwegian-English translation memories that have been collected through the ELRC, very few are between Nynorsk and English. As for Bokmål and Nynorsk, there are

[24] http://espeak.sourceforge.net
[25] Originally produced by the Norwegian company Max Manus which was recently bought by Omilon.
[26] https://www.nynodata.no/norsok
[27] https://pypi.org/project/summa/
[28] https://prokom.no/kari/
[29] https://www.nrk.no/vestland/kommune-robot-klarer-ikke-svare-pa-enkle-sporsmal-1.14191246
[30] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-60/

several corpora containing government documents and official web pages that exist in both Bokmål and Nynorsk. Some Nynorsk-Bokmål parallel corpora have been created on the basis of this, such as Målfrid,[31] as well as a few translation memories from Nynorsk to English. The most prominent Nynorsk-Bokmål corpus is the manually corrected output of the Nynorsk press agency Nynorsk Pressekontor's Apertium-based pipeline.

Multimodal corpora

While the NDC contains recordings and transcriptions done between 2006 and 2012, i. e., fairly recently, the Language Infrastructure made Accessible (LIA) corpus contains older dialect recordings. Whereas the transcriptions in the NDC are normalised to Bokmål, the transcriptions in LIA are normalised to Nynorsk. All in all, it contains a comparable amount of audio and text data to the NDC, but has no video data.

The previously mentioned Norwegian Parliament Speech Corpus contains an equal amount of orthographically transcribed parliament procedures for Nynorsk as for Bokmål, since all the texts were translated, corrected and made available in both written forms.

Lexical/conceptual resources

The most important lexical resource for Nynorsk is its version of the aforementioned Norwegian Word Bank, which is a lexical database for Norwegian Nynorsk reflecting the official standard orthography as defined in the Nynorsk dictionary *Nynorskordboka*.[32] Similar to the Bokmål dictionary, the Nynorsk dictionary is jointly owned by the Norwegian Language Council and the University of Bergen, where it is maintained and hosted. Both resources are freely available for download and use in language technology. The word bank is continuously updated by the Language Collections at the University of Bergen, and the database reflects the Nynorsk norm from 2012.

Nynorsk termbases are scarce. While some domain-specific termbases exist for Bokmål, very few terms appear in their Nynorsk parallel, for instance in the national terminology portal *Termportalen*.[33] Exceptions include the already mentioned SNORRE termbase and to some extent the EU termbase from the Ministry of Foreign Affairs as well as a few others.

Nynorsk pronunciation lexicons for speech technologies are equally scarce. The Lingit pronunciation lexicon for Nynorsk was developed for TTS voices, and contains 570,390 lexical units consisting of a morphologically inflected word form, an X-SAMPA phonemic transcription, morphosyntactic features and the lemma.

The NLB pronunciation lexicon for Bokmål contains a file with 352,788 automatically generated Nynorsk transcriptions based on the Norwegian Word Bank for Nynorsk.

Models and grammars

While there is no distinct transformer language model for Nynorsk, the two Norwegian BERT-based models NB-BERT and NorBERT have been trained on data containing Nynorsk text. In principle, it is possible to fine-tune one of these models on Nynorsk-only annotated data. However, this depends on the availability of annotated Nynorsk data for a given linguistic analysis, and such data is lacking.

[31] Målfrid contains roughly 350M Nynorsk tokens
[32] https://ordbokene.no
[33] https://term.uib.no

Tools and Services

The availability of online dictionaries in Norwegian is quite good for both Bokmål and Nynorsk. In addition to the previously mentioned *Nynorskordboka*, the *Norsk Ordbok* dictionary provides access to dialectal and Nynorsk-specific words,[34] with lexical and grammatical as well as etymological information.

Due to the similarities between Nynorsk and Bokmål, machine translation between the two languages yields fairly good results, and the translation service at the Apertium platform performs very well in this language pair.

As for translations between Nynorsk and English, Nynorsk and Bokmål are not separate language options on Google Translate, and both are translated under the "Norwegian" umbrella. Because of the predominance of Norwegian Bokmål in the digital sphere, the translated text will always be rendered in this variety.

There are very few translation memories between Nynorsk and English in the corpus gathered for the ELRC, and not enough data to include Nynorsk as an option in eTranslation. With the exception of the direct translation developed by the PRINCIPLE project, translations between Nynorsk and English will use Bokmål as a pivot language.

There are several coding libraries that provide functionality for automatic linguistic analyses of text. These libraries have often trained their machine learning models on the training material found in the Universal Dependencies (UD) repository.[35] UD for Norwegian is a conversion of the NDT, and some libraries, e. g., Stanza,[36] offer support for Nynorsk and Bokmål. Others which depend on robust word embeddings do not, e. g., SpaCy. This is due to the lack of enough data to train such embeddings specifically for Nynorsk.

OBT+Stat, i. e., the extended OBT with statistical disambiguation for tagging, was mentioned in Section 4.1. While the statistical component is only available for Bokmål, the original POS-tagger component, OBT, can tag Nynorsk text.

Until now, speech processing tools have been almost non-existent for Nynorsk. Nuance has just announced that they will provide speech-to-text for Norwegian Nynorsk as well as Bokmål. Neither Microsoft nor Google provide speech synthesis or speech recognition for Nynorsk. As an exception, the Norwegian Library of Talking Books and Braille provides audio material which has been automatically generated from Nynorsk text with their synthetic voice Hulda. Some commercial writing tools for children also offer text-to-speech for Nynorsk.

Stortinget (the Norwegian Parliament) is in the process of implementing automatic transcription and further automatic minuting for their meetings. Another summarization service that exists is Oppsummert by the newspaper *Aftenposten*. While this most likely is something the journalists do manually, the effort can be used as a dataset for training automatic summarization models.

4.3 Comparison between Norwegian Nynorsk and Bokmål

Large web-based corpora have an unequal distribution between the two languages, reflecting not only the relatively low percentage of Nynorsk texts, but also the tendency towards developing language technologies for Bokmål only. Some rough estimations have been made from the NLN's corpora suggest that 5-10% of their texts are in Nynorsk. The Målfrid corpus contains just under 10% Nynorsk text.

While some resources, such as the NDT and the Norwegian Parliament Speech Corpus, have an equal distribution of Bokmål and Nynorsk, multilingual LT service providers tend

[34] http://no2014.uib.no/perl/ordbok/no2014.cgi
[35] https://github.com/UniversalDependencies
[36] https://stanfordnlp.github.io/stanza/available_models.html

to only develop a Bokmål version for Norwegian, rather than both Bokmål and Nynorsk versions. The fact that relevant language resources (LRs) are predominantly available in Bokmål, or that the demand for Nynorsk versions of the service might seem disproportionately low, could be contributing factors to this situation.

Many documents and public-facing text produced by public institutions and organisations exist in both Bokmål and Nynorsk, such as can be found in the Målfrid corpus. Translation memories between the written standards, such as the one from the aforementioned press agency Nynorsk Pressekontor, are useful for developing LT that is compatible for both.

A Norwegian peculiarity is the variation in information that is rendered by search engines, depending on whether the search is done in Bokmål or Nynorsk. A Norwegian web page is more likely to be rendered as a search result when it is written in Bokmål, because the search language is more likely to be Bokmål. The algorithms on the big platforms do not seem to cater to the parallelism between the two languages. While many words are similar and will render the same results, words that have a slight difference in spelling will only render results in one language, unless the lexicon and /or terminology has been parallelised in the search engine. Because public bodies are required to publish a text in one of the two languages, but not the same text in both languages, even documents such as laws and regulations "disappear" in the result from their internal search engines. Therefore, many public bodies render different information to Bokmål users and Nynorsk users. Nynorsk users will receive less information if the request is made in their language. Ironically, this may be just as big a problem for a Bokmål user as for a Nynorsk user, since the latter may be more conscious of the presence of the majority language in the public domain and hence of the need to perform searches in both languages.

4.4 Projects, Initiatives, Stakeholders

National Programmes for LT

Even though the last few years have seen a tremendous focus on artificial intelligence, the importance of language technology as an important component in AI programmes has only been recognised sporadically. Norway has no research programme specifically directed towards language technology. There is, however, a national programme for AI, funded by the Research Council of Norway, where LT projects may apply (Collaborative Project on Digital Security and Artificial Intelligence, Robotics and Autonomous Systems).

Research infrastructures

The three largest repositories for Norwegian language data are the Language Bank, the text laboratory at the University of Oslo and the depot hosted by the CLARINO Bergen Centre. The latter includes INESS, which is an open platform for building, accessing and visualizing treebanks. All three repositories take part in the Common Language Resources and Technology Infrastructure Norway (CLARINO). Out of these, only the Language Bank has adopted a policy of complete openness. With few exceptions, the resources are free of any restrictions, and they are licensed as CC-0. In the other repositories, many resources have a relatively free license, and often it is only required that the origin of the resource be cited (CC-BY).

Research projects/initiatives

Many of the gaps in Norwegian language technology and language resources that have been identified are already being addressed by ongoing projects. All major universities in Norway conduct research on language technology and/or AI. Among the most recent projects

is NorwAI,[37] jointly funded by the Research Council of Norway and the project partners. It aims at developing language technologies for Scandinavian languages, including conversational search in natural language. It also seeks to provide solutions to the scarcity of domain-specific resources through transfer-learning methods.

Another project is SCRIBE, which seeks to develop an advanced speech-to-text transcription system for spontaneous speech. SCRIBE is also jointly funded by the Research Council of Norway, and the research partners are both public institutions and private companies.

The goal of the ongoing SANT (Sentiment Analysis for Norwegian Text), coordinated by the Language Technology Group at the University of Oslo, is to create open resources for sentiment analysis for Norwegian. The public broadcasting corporation NRK and two private media groups contribute to the project.

The Målfrid project collects all available digital texts from the public sector in Norway. An effort like this will ensure the availability of unstructured text data of a more recent date, but no continual addition to annotated datasets like NDT is in place. Similar to Målfrid, there is an ongoing effort to secure continual delivery of newspaper text between the Language Bank and the media organizations in the newly established centre for research-driven innovation MediaFutures.[38] MediaFutures has a working group dedicated to language technology, where one of its goals is to develop both a data set as well as a system for event extraction from text.

CLEANUP is a project funded by the Research Council of Norway and run by the Norwegian Computing Center, which aims to develop tools and techniques to automatically anonymise unstructured text data from an array of domains. The project partners include the University of Oslo, NTNU, Universitat Rovira i Virgili (Spain), the National Archives of Norway and the Norwegian Labour and Welfare Administration, as well as partners from the private sector such as Lovdata, Gjensidige and DNB. Such a consortium envisages not only the will to invest in new technology, but also the ambition of putting it in production once it reaches a mature state.

The project Universal Natural Language Understanding, financed by the Research Council of Norway, builds upon the UD standard for syntactic treebanks. The goal of the project is to convert the syntactic representation to machine-readable semantic representation. The ambition of the project is a general conversion procedure for all of the 90 languages found in UD.

LT providers

As LT has become an integrated part of all aspects of society, it becomes harder and harder to pinpoint providers working with LT only. However, the start-up and SME ecosystem in Norway has fostered some companies with LT as a core of their business model, including chatbots (Kindly and BoostAI), writing support tools (Lingit), speech technologies (Max Manus), translation technologies (Semantix and NTB Arkitekst), and argumentative structures in text (Disputas), to name some examples.

5 Cross-Language Comparison

The LT field[39] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded re-

[37] https://www.ntnu.edu/norwai/
[38] https://mediafutures.no/norwegian-language-technologies/
[39] This section has been provided by the editors.

sults unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[40] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing (e. g., facial expression recognition)
  - Human-computer interaction (e. g., tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

5.2 Levels of Technology Support

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

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

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

3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[41]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[42] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[43]

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

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

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 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

[41] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e., 3%, 10% and 30%) were then used to define the bands per application area and resource type.

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

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

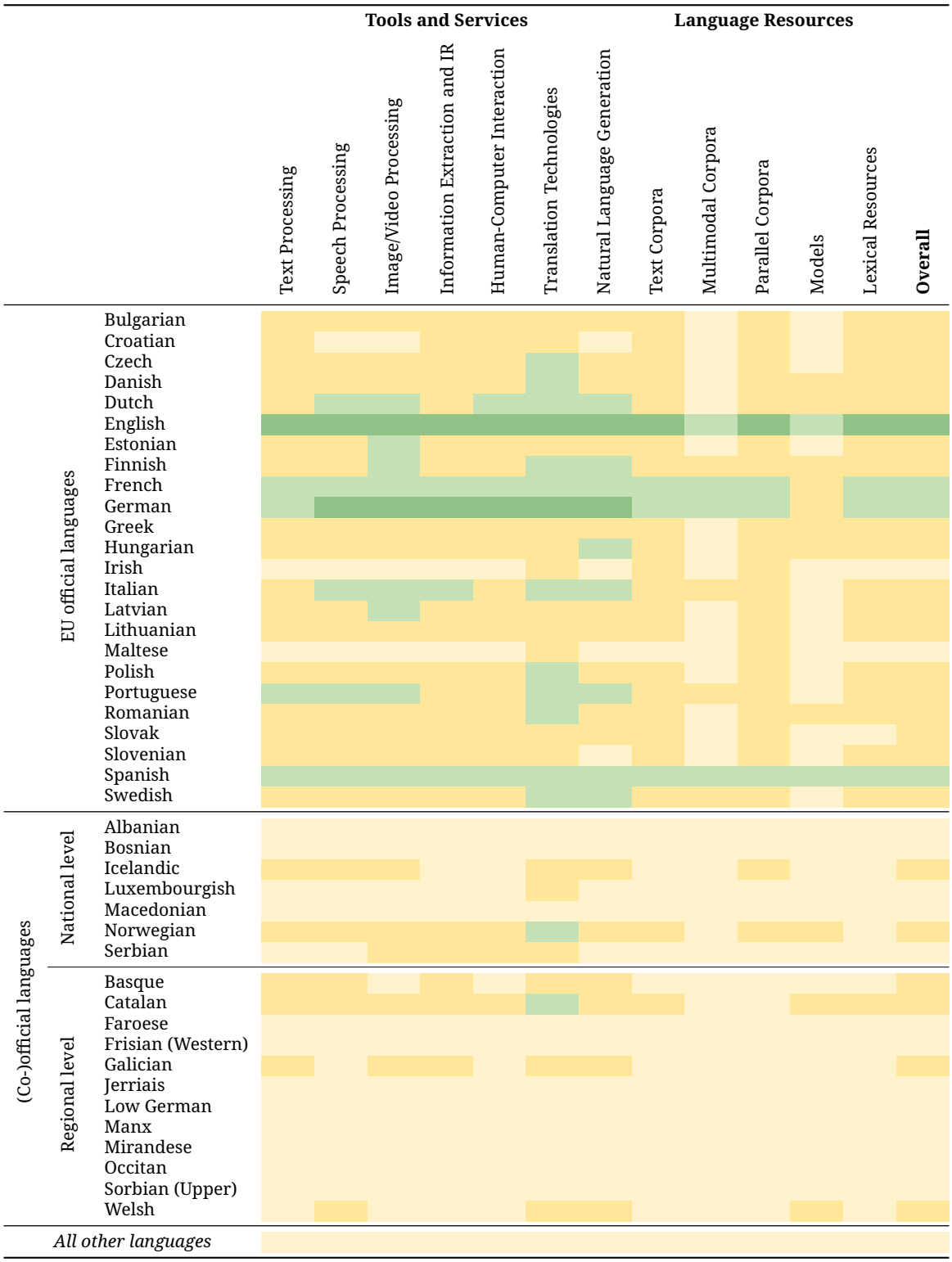

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

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,[44] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

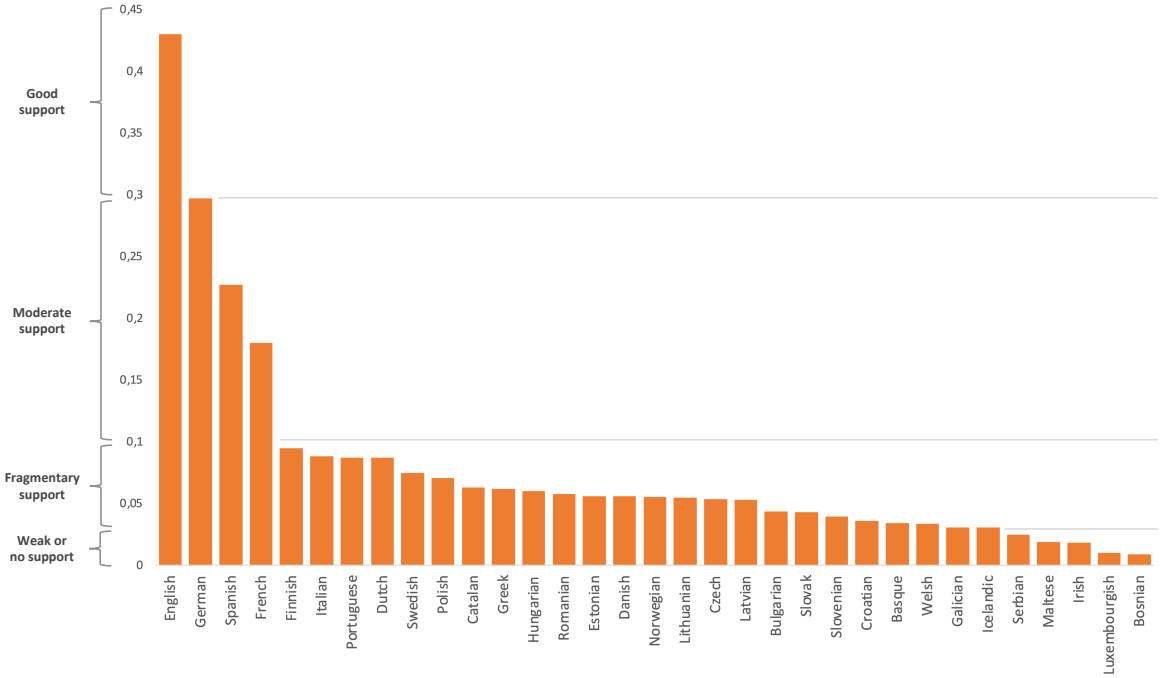

Figure 1: Overall state of technology support for selected European languages (2022)

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, 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 im-

[44] 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.

balance, 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

The great interest in AI in both the public and private sectors in combination with a digitally competent population and a well-developed digital infrastructure is definitely a reason to be optimistic about the future of Norwegian language technology. The recent reports on language, digitalisation and artificial intelligence show increased consciousness of the importance of language technology as a component in AI, and the importance of data sharing for language technology.

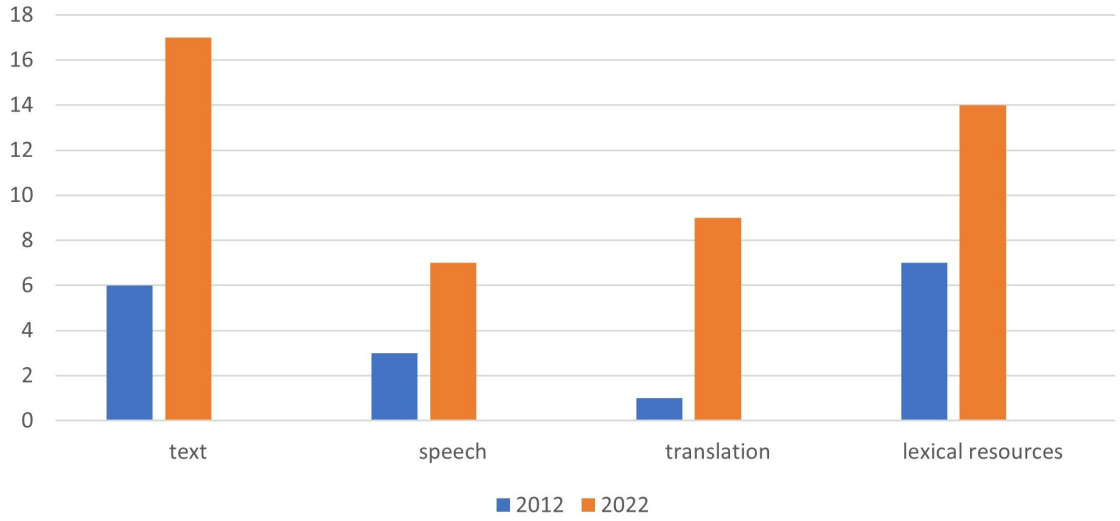

Figure 2: Number of language resources in the Language Bank in the years 2012 and 2022

Compared to the situation in 2012, as described in the first META-NET White Paper (Smedt et al., 2012), the amount of Norwegian language resources has increased substantially, and there is a political will to finance these basic resources. This increase does not show in the European comparison in chapter 5, since the other European languages have had a similar

increase in their resources. To illustrate the Norwegian increase, we include figure 2, which shows the number of resources in the Norwegian Language Bank in 2012 compared to 2022.

As for the quality of Norwegian language technology, no overreaching assessment has been made of the improvement we assume has taken place. Given the Norwegian language situation, it would be interesting to compare not only the Norwegian scores with the scores of other languages, but also the Bokmål and Nynorsk varieties as well as speech recognition for dialects that differ from the standard Eastern Norwegian variety.

As for the Norwegian language in the digital sphere, there is no sign of digital extinction for Bokmål. Bokmål is well supported by digital platforms, both nationally and internationally. The situation is not equally unproblematic for Nynorsk. Nynorsk is less used in the digital sphere and may find itself caught in the vicious circle of minority languages, where the scarcity of language data leads to lower-quality language technology, thus making the majority language more attractive from the point of view of the digital user. This attractiveness will in turn provide less Nynorsk data to be used in new language technology, etc. Due to the lack of Nynorsk in the digital sphere and modern language technology's preference for big data, it must be a priority for decision makers to strengthen LT for the lesser used language to avoid weakening its equal status. While there are certain profitable synergies when developing parallel language technologies for both languages, there is also a need for parallel development of basic resources.

To ensure the vitality of Norwegian in the future, the new Norwegian Language Act of 2022 states that the Norwegian institutions on the state level have a special responsibility to maintain and develop Norwegian as a language that can be used in all circumstances and in all domains, with Nynorsk as a particular responsibility. One way to live up to this responsibility is for the public sector to procure language technology that takes the particular Norwegian language situation into account. When ordering new products, they must make sure the language technology supports both languages.

A best practice example of a public institution assuming responsibility for their Norwegian language data is demonstrated by the Knowledge and Documentation Department in the administration of the Norwegian Parliament, which is responsible for the official reports from the parliamentary sessions. With the aim of introducing automatic transcriptions of the sessions, they have ensured that the data deriving from the transcriptions are handed over to the Language Bank at the National Library. These datasets are particularly important for Norwegian speech technology because of dialectal variation in parliamentary speech and because the politicians themselves decide whether they want their speeches written down in Bokmål or Nynorsk. There is reason to believe that other parts of the public sector have more data, than they are aware of, that can be made available for use in AI and LT.

While a great effort has been made to gather new resources, we are not there yet. Ongoing projects such as NorwAI, Scribe and SANT aim to fill some of the most urgent gaps in language technology tools and resources, providing contact between research environments and private companies. Although no research programme is specifically directed towards language technology, LT falls within the scope of more general programmes such as ICT and digital innovation, AI programmes and infrastructure programmes. According to section 1 of the new language act, public bodies, such as the Research Council of Norway, have a special responsibility for promoting Nynorsk as the least used written Norwegian language. The extra cost of developing parallel versions of Bokmål and Nynorsk technologies should be taken into consideration when funding future LT research programmes. They are two separate languages for which we need two sets of language technology, and both languages require financing.

To summarise: these are the main issues that need to be resolved if Bokmål and Nynorsk are to thrive in the digital sphere in the future:

- Continue the creation of missing tools and resources, among others, more text data for

Nynorsk, more domain-specific data, lexical/terminological resources, in particular for Nynorsk, as well as speech data that cover dialects and Nynorsk and tools for semantic parsing.

- Such resources and tools should be made available under permissive licences, in order to ensure their reusability.
- Datasets that allow text analysis above the sentence level, such as a coreference corpus.
- Continue to raise awareness of the importance of language data.
- Continued participation in international research projects and other projects that focus on language technology, such as ELE, ELG and ELRC.
- Ensure sufficient funding for language-specific LT for Bokmål and Nynorsk.
- Public sectors must take on their new responsibility as required in the new language act and ensure parallel versions of Norwegian language technology in public procurement. Develop standard formulations for public procurements to give the public sector the rights to language resources which emerge from translations and other services.
- Downstream (user-driven) quality assessment of Norwegian language technology tools and services in order to compare the quality of Nynorsk and Bokmål tools and services as well as dialect understanding.

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.

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. NorDial: A preliminary corpus of written Norwegian dialect use. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 445–451, Reykjavik, Iceland (Online), May 31–2 June 2021. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/2021.nodalida-main.51.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Helge Dyvik, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Martha Thunes. NorGramBank: A 'deep' treebank for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3555–3562, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1565.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online), 2021. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/2021.nodalida-main.3.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. Large-scale contextualised language modelling for norwegian, 2021.

Language Council of Norway. Språkstatus 2021, December 2021. URL https://www.sprakradet.no/Vi-og-vart/Publikasjoner/sprakstatus1/sprakstatus-2021/. Accessed 18 January 2022.

Ministry of Culture. Prop. 108 L. Proposisjon til Stortinget (forslag til lovvedtak) Lov om Språk (språklova), 2020. URL https://www.regjeringen.no/no/dokumenter/prop.-108-l-20192020/id2701451/. Accessed 14 January 2022.

Ministry of Local Government and Regional Development. Én digital offentlig sektor: Digitaliseringsstrategi for offentlig sektor 2019-2025, 2019. URL https://www.regjeringen.no/no/dokumenter/en-digital-offentlig-sektor/id2653874/. Accessed 18 January 2022.

Ministry of Local Government and Regional Development. Nasjonal strategi for kunstig intelligens, January 2020. URL https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/. Accessed 18 January 2022.

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.

Koenraad De Smedt, Gunn Inger Lyse, Anje Müller Gjesdal, and Gyri S. Losnegaard. *Norsk i den digitale tidsalderen (bokmålsversjon) – The Norwegian Language in the Digital Age (Bokmål Version)*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 9 2012. URL http://www.meta-net.eu/whitepapers/volumes/norwegian-bokmaal. Georg Rehm and Hans Uszkoreit (series editors).

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. The norwegian dependency treebank. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 789–795, 2014. European Language Resources Association, ISBN 978-2-9517408-8-4.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.