EUROPEAN LANGUAGE EQUALITY

D1.9

Report on the Danish Language

Authors	Bolette Sandford Pedersen, Sussi Olsen, Lina Henriksen
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	D1.9
Deliverable title	Report on the Danish Language
Type	Report
Number of pages	26
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	Bolette Sandford Pedersen, Sussi Olsen, Lina Henriksen
Reviewers	Stefanie Hegele, Tea Vojtěchová
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
ASR	Automatic Speech Recognition
CLARIN	Common Language Resources and Technology Infrastructure
CLARIN-DK	CLARIN Denmark
COR	Central Word Register for Danish
DLE	Digital Language Equality
DTU	Technical University of Denmark
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
GPS	Global Positioning System
GPU	Graphics Processing Unit
HPC	High-Performance Computing
LR	Language Resources/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
NLG	Natural Language Generation
NLP	Natural Language Processing
SME	Small and medium-sized enterprise
SR	Speaker Recognition
UD-DDT	Danish Universal Dependencies Treebank
UCPH	University of Copenhagen

Abstract

Today, language-centric AI is being introduced in nearly all aspects of the Danish society. We use chatbots when we communicate with the municipality and our GPS speaks Danish when guiding us through the city. We apply machine translation when communicating across borders, we monitor the attitude towards our brand on social media, and deep learning algorithms help us make important decisions on health and social welfare. Some people even speak to a robot in their private homes in order to control their heating system, their electricity, and their choice of music and television. As a natural consequence of this development, the importance of high-quality Danish language resources and technology (henceforth LT) is becoming more and more broadly acknowledged at all levels. Merely transferring technologies from English without adapting smoothly to the Danish language and culture most often results in poor systems which are not fully functional and furthermore not inclusive to all parts of the society.

Several factors play a role in how fast and how well a language community adapts to new technological advances. Even if Denmark is one of the most digitised countries in the world, the investments in Danish LT have been somewhat delayed by factors such as the country's relatively small size, both as a language community and as a commercial market, together with our high proficiency of English. Specific characteristics of the Danish language may also play a role, e. g., Danish speech technology is challenged by the tendency to perform phonetic reduction in spoken Danish, by our large number of vowels, by our famous glottal stop etc.

In this report we give an up-to-date overview of the current level of Danish LT including services such as machine translation, virtual assistants, speech systems, sentiment analysis, automatic abstracting, fake news detectors etc. Even if such Danish LT services are now at hand and often also freely available to the public, their *quality* still needs to be improved in order to make them really useful for the Danish users and actually comparable with similar services for English. To this end, large, high-quality language resources and data sets still prove to be the real bottleneck.

Several new LT players are in recent years entering the Danish scene, and there is an increased awareness of sharing and reusing language resources and data sets across public institutions, academia and industry. This is really good and promising news.

Furthermore, new, large governmental initiatives within the area of AI and LT are currently being embarked. In 2019, the Danish Government adopted a comprehensive AI strategy which includes an element of LT with special focus on *language understanding* and *speech technologies*. Following this line, new projects have recently been launched with the aim of compiling the Danish resources necessary for the development in these areas.

This being said, the need for a continuous coordinated effort now and in the future, based in both governmental initiatives, industry, public research and education, still remains. Such continuous efforts are mandatory in order to keep Danish on track towards a digitally fully functional language whilst also providing future language-centric AI solutions.

Dansk resumé

Kunstig intelligens (eller Artifical Intelligence – AI) vinder i disse år indpas på snart sagt alle niveauer i det danske samfund. Vi bruger chatbots når vi taler med kommunen, vores gps taler dansk når den leder os gennem byen, vi bruger maskinoversættelse når vi skal kommunikere på tværs af landegrænser, og algoritmer hjælper os med at tage vigtige beslutninger inden for sundhed og velfærd. Nogen taler ligefrem med en robot i deres private hjem når de skal skrue ned for varmen, tænde for lyset eller vælge musik- og fjernsynsprogrammer. Efterhånden som teknologien på denne måde kommer tættere på vores almindelige liv som

medborgere og privatpersoner, får sprog og kultur en større betydning for den teknologi der udvikles. Vi taler i den forbindelse om *sprogteknologi* og *sprogcentreret AI*.

Denne drejning mod sprogcentreret AI øger fokus på betydningen af *danske sprogressourcer*, som netop danner baggrund for udvikling af sprogteknologi af høj kvalitet. Altså tekst- og lydsamlinger, ordbaser og sprogmodeller der dækker det danske sprog på en nuanceret måde. For det står også mere og mere klart at direkte overførsel af engelsk sprogteknologi til dansk uden tilstrækkelig tilpasning til det danske sprog og samfund, resulterer i halvdårlige løsninger som ikke er fuldt funktionelle, og som slet ikke er inkluderende overfor alle dele af samfundet. Det kan med andre ord blive en udfordring at skabe bred opbakning og tiltro til ny teknologi hvis ikke dansk sprogteknologi har en høj kvalitet.

Der er flere faktorer der har betydning for hvor hurtigt og hvor smidigt et sprogsamfund tilpasser sig de nye teknologiske muligheder. Selv om Danmark er et af de mest digitaliserede lande i verden, så betyder vores lidenhed – både som sprogsamfund og som kommercielt marked – at dansk sprogteknologi er udfordret. Det faktum at vi er relativt gode til engelsk, har givetvis også spillet en rolle for hvor hurtigt vi har sat ind på at udvikle teknologi på dansk. Derudover er der nogle egenskaber ved det danske sprog som muligvis har haft en vis forhalende effekt. Fx har dansk taleteknologi været udfordret af den udbredte brug af fonetisk reduktion, dvs. at vi trækker mange lyde sammen når vi taler, sådan at ordgrænser er svære at lokalisere automatisk. Vores mange vokallyde i dansk og brugen af stød som betydningsadskillende træk (som i *ham* (i modsætning til *hende*) uden stød vs. *ham* (som i *slangeham*) med stød) er også karakteristika der har givet dansk taleteknologi en svær start.

I denne rapport giver vi et billede af hvor dansk sprogteknologi står lige nu, dels i forhold til grundlæggende sprogressourcer som sprogmodeller og ordbeskrivelser, dels i forhold sprogtjenester som maskinoversættelsessystemer, virtuelle assistenter, talesystemer mv. Selv om sådanne danske sprogtjenester nu er tilgængelige i større eller mindre grad, så er der stadig et stykke vej før de alle er fuldt praktisk anvendelige og i øvrigt sammenlignelige med sådanne sprogtjenester for engelsk. Det der mangler for at vi kommer på niveau, er i høj grad danske sprogressourcer og datasæt som er tilstrækkelig store og af tilstrækkelig høj kvalitet. De udgør så at sige flaskehalsen for den videre udvikling. Som rapporten viser, er der dog sket rigtig meget bare de seneste 2-3 år især drevet af paradigmeskiftet over imod at anvende *maskinlæring* frem for regelbaserede teknikker. For det første er der kommet en del nye sprogteknologiske spillere på banen, og med den udvikling er opmærksomheden omkring *deling af sprogressourcer* på tværs af forskning, offentlige institutioner og det private erhvervsliv øget betragteligt. Dette er en virkelig god nyhed som kan være med til at booste dansk sprogteknologi og gøre at vi tager et gevaldigt spring fremad inden for kort tid!

Hertil kommer at flere nationale initiativer er sat i gang i de senere år til netop at understøtte produktionen af sprogressourcer, herunder regeringens AI-strategi fra 2019. Strategien indbefatter en satsning på sprogteknologi med særligt fokus på nyudvikling inden for *taleteknologi* og *sprogforståelse*. Konkret betyder det i første omgang at der er etableret en national sprogportal til deling af sprogteknologiske komponenter (https://sprogteknologi.dk), og at der igangsættes udviklingsprojekter inden for de prioriterede områder.

Når dette er sagt, så er behovet for et koordineret og kontinuerligt fokus på udvikling af dansk sprogteknologi af stor vigtighed, også fremover. Hvis vi ikke vil lade udenlandske techgiganter afgøre hvornår og hvor godt vores maskiner skal tale dansk, er det nødvendigt at forskning, offentlige institutioner og private aktører står sammen og fortsat udvikler og deler de komponenter som der er behov for. Dette er nødvendigt hvis dansk skal forblive et fuldt funktionelt sprog også i fremtidige AI-løsninger.

1 Introduction

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

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

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 The Danish Language in the Digital Age

2.1 General Facts

Danish: A Mainland Scandinavian Language

Danish is the official language of Denmark, which has approx. 5.831 million inhabitants.[1] Approx. 90 percent have Danish as their mother tongue. Danish is also the native or cultural language of around 50,000 Germano-Danish citizens living in the south of Schleswig. In the Faroe Islands and Greenland, the law of autonomy guarantees official equality of Danish alongside the Faeroese and Greenlandic languages, and Danish is an obligatory subject in schools.

Danish is a North Germanic language and derives from the East Norse dialect group. Modern spoken Danish, however, is classified as a Mainland Scandinavian language group together with Norwegian and Swedish. This more recent classification is based primarily on the mutual intelligibility among these three languages.

Danish phonology distinguishes from several of its neighbour languages by exhibiting for instance a very large number of vowels and by having glottal stop as a meaning differentiating feature (‘hund’ (‘dog’) with a glottal stop vs ‘hun’ (‘she’) with no glottal stop). Further, phonetical reductions are very common, in particular among young people, a fact which complicates Danish speech technology since for instance word boundaries become hard to identify.

Written Danish applies the Latin alphabet with three extra letters, namely æ, ø, and å. Still today, Danish users encounter problems in several technical services that have not included these extra letters in a seamless fashion.

[1] Cf. among others https://denstoredanske.lex.dk/dansk, https://da.wikipedia.org/wiki/Dansk_(sprog), https://dsn.dk, https://sproget.dk. See also the META-NET White Paper on the Danish language in the digital age (Pedersen et al., 2012).

In written language, the fact that compounds are spelled as one word (as in other Germanic languages) complicates the construction of language tools, and further, compounds are generated very dynamically and therefore only partially accounted for in dictionaries.

Also the very extensive use of particles with semi-lexicalised meanings poses a challenge to both the syntactic and semantic analysis of LT systems. The constructions often occur discontinuously in spoken and written Danish, as in 'du skal læse den lange, indviklede tekst op for publikum' (lit: 'you should read the long and complicated text up (out loud) to the audience'), a fact which tends to require large amounts of language data in order to be well represented in the corresponding language models.

What further complicates automated analysis is the heavy use of topicalisation and movement of complements, a phenomenon that Danish shares with the other Scandinavian languages, as seen in for instance: 'Dette argument ved vi godt hvem der har fremlagt', ('This argument we know very well who has presented').

Foreign Influence on Danish

The influence of the English language on Danish language users is increasing (for a recent account, see Gottlieb, 2020). This is seen in education where a substantial number of courses at Danish universities are taught in English with the consequence that several – in particular technical – domains are more or less lacking Danish terminology.

Likewise, in industrial settings, English is more and more often chosen as the company language, also meaning here that terminology is mainly being developed in English.

The English influence via social media is also evident, and all in all these developments entail that new loan words and fixed phrases are taken in from English with increasing speed – often in cases where perfectly equivalent Danish expressions exist.

Loan words and fixed phrases do not influence the *language system* as such, however, syntax is also influenced in some particular cases. For instance, some Danish verbs change valency pattern because of the influence from English, as is the case for 'at gro' ('to grow') which is originally an intransitive verb in Danish (but transitive in English) and which is now beginning to occur as transitive, as in 'kan man gro trøfler i Danmark?' ('can you grow truffles in Denmark?'). In addition, word order, including the placements of adverbials, tends to be increasingly influenced by English.

2.2 Danish in the Digital Sphere

97% of the Danish population aged 12 years and above have access to the Internet, and 91% use the Internet on a daily basis (2021).[2] The Internet country code top-level domain from Denmark is the .dk domain. More than 1,375,000[3] websites have this domain, and almost all of these are in Danish.

Apart from monitoring the number of websites written in Danish, a complete overview of the Danish-English distribution in relation to online communication, e. g. social media is not directly available. Statistics Denmark publishes a yearly report on the use of online communication in Denmark,[4] and this report maps out important aspects of Danes' access to and use of online solutions including the use of e-commerce and social media. The use of Danish vs English is however not among the measured aspects; Danes are generally excellent

[2] Statistics Denmark https://www.dst.dk/da/Statistik/emner/kultur-og-fritid/digital-adfaerd-og-kulturvaner/digital-adfaerd

[3] DK-hostmaster: https://stats.dk-hostmaster.dk/domains/total_domains/yearly

[4] https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/it-anvendelse-i-befolkningen

English speakers[5] and therefore use English often and with little effort. However, in spite of this high proficiency of English, the vast majority, 80%, of the Danish population feel insecure about shopping online in another language than Danish.[6]

3 What is Language Technology?

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

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

[5] Education First's English Proficiency Index: https://www.ef-danmark.dk/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2021/ef-epi-2021-english.pdf

[6] https://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=29450&sid=itbef2020, p. 20

[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).

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]

[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).

4 Language Technology for Danish

Research and developement within language technology has been ongoing in Denmark for several decades. Where development of tools and resources was relatively sparse in the early years and mostly had the character of prototypes, language technology is now – even if we do not necessarily notice it – an integrated facility in more or less all aspects of the digital society, from search engines to virtual assistants, to translation services and chatbots.

This being said – and even if Denmark has been and still is one of the most digitised countries in the world – the understanding of the language dimension in technology and likewise seeing the collection and enrichment of language data as a joint commitment is only just recently beginning to resonate in the Danish society. The thing is, that even if Danish LT services are now at hand in most areas of technology, the *quality* still needs to be improved in order to make these services really useful for the Danish users so that they are actually comparable with similar services for English. To this end, large, high-quality language resources prove to be the real bottleneck.

In the following sections, a broad overview of recent and current developments is given; not only with respect to these grounding resources, but also concerning the development of language services developed from such resources. The overview is primarily based on a number of existing LT repositories for Danish.[9]

4.1 Language Data

Text Corpora

Large text collections are basic requisites for most developments in LT. Text corpora of general language have typically been collected by institutions that develop dictionaries, such as The Danish Language Council and the Society for Danish Language and Literature. These institutions host very large well-balanced corpora today, but due to property rights they are not for the entire part open source and ready to use for industry. For research and non commercial purposes, the DK-CLARIN Reference Corpus of General Danish of 45 million words has been available for a decade at the CLARIN-DK repository.

To build high-quality and representative language models, however, more data is required, preferably open source. This recently led to the development of the Danish GigaWord initiative, a freely available billion word corpus of Danish texts assembled among a large group of researchers (Strømberg-Derczynski et al., 2021).[10] In addition, a large number of smaller, more domain specific and historical corpora can be found, for instance, on the resource portal sprogteknologi.dk and at clarin.dk.

Corpora that are human-annotated with linguistic and other information types are of particular value as gold standards for supervised learning. Even if such annotated corpora are still needed for many areas of Danish, some freely available corpora of varying size do exist with the annotation of part of speech, word senses, syntactic structure, sentiment etc. In addition, a human-annotated data collection for Danish summarisation has been created recently (Varab and Schluter, 2020), see also Pauli et al. (2021) for a collection of newly compiled datasets at DaNLP.

[9] Examples of LT repositories that have been accessed: the Danish CLARIN platform, CLARIN-DK, http://clarin.dk, the repository of The Danish Agency for Digitisation, http://sprogteknologi.dk, COASTAL's repository, https://coastalcph.github.io, (University of Copenhagen), Awesome Danish at the Danmarks Tekniske Universitet (DTU), https://github.com/fnielsen/awesome-danish, the Alexandra Institute's repository, https://github.com/alexandrainst/danlp, Centre for Language Technology's (University of Copenhagen) repository, https://github.com/kuhumcst, Centre for HUmanities Computing at Aarhus University, https://github.com/centre-for-humanities-computing/DaCy, the Society for Language and Literature's list, https://korpus.dsl.dk/resources/, and ITU's github, https://github.com/ITUnlp, and Stanford NLP https://github.com/stanfordnlp/stanza/.

[10] Note that GigaWord subsumes all freely available corpora from CLARIN-DK.

Language models

Several statistical and neural language models have been processed for Danish in recent years and are based primarily on the above mentioned text collections. Schneidermann et al. (2020) reports on six different models (trained with either word2vec or fasttext) with different correlations with a hand-crafted similarity data set. Out of the six models, a word2vec model processed by the Society for Danish Language and Literature (Sørensen and Nimb, 2018) was the one that correlated the best with the hand-crafted similarity dataset according to Schneidermann, probably because they had a larger and better balanced corpus at hand.

Just recently, also a number of contextualised, pretrained models have been processed for Danish.[11] Overall, these contextualised models enable improved language processing with for instance a better grasp of the variation of word meaning in running text. The Scandival benchmark[12] evaluates these and other models for the Scandinavian languages and benchmarks them according to different tasks.

Multimodal, Parallel, and Speech Corpora

Multimodal corpora where video recordings are transcribed and annotated also exist for Danish at a smaller scale. An example is the NOMCO corpus (Paggio and Navarretta, 2017), an annotated multimodal collection of conversational Danish which annotates Danish video conversations with gestures. Such corpora can serve as a basis to model how gesture and non-verbal behaviour contribute to communication.

Parallel text corpora are primarily used to build statistical models for machine translation, and these models are highly dependent on really large amounts of text data within all domains. The number of parallel corpora including Danish has increased somewhat over the last few years; especially corpora where one language is English and the other is Danish. Other bilingual corpora with Danish as one of the languages are rather scarce. Most of the accessible multilingual corpora have English as the source language which means that all other languages in the corpus are primarily parallel to English. Multilingual corpora available are mostly collected from EU institutions and via webcrawls. Overall, there is a lack of parallel text corpora for Danish in combination with languages other than English (see Kirchmeier et al., 2019, for an account on this). In recent years, however, the EU initiative European Language Resource Coordination (ELRC)[13] has helped increase awareness on the value of parallel corpora,in collaboration with three nationally located anchor points. To this end, both public institutions as well as a few private companies (like the translation company Semantix) have contributed with the donation of parallel text resources including Danish.

Large public speech corpora are generally in short supply for Danish, a fact which complicates the development of speech technologies for Danish. However, few such resources exist at a medium scale, namely the Danish NST ASR Database available at the Norwegian Språkbanken and compiled originally by the company Nordisk Sprogteknologi (NST); DanPASS compiled at the University of Copenhagen (Grønnum, 2006); and the Danish Parliament Speech Corpora (Hansen and Navarretta, 2018; Kirkedal et al., 2020). The production of a large, transcribed and time-encoded speech corpus is foreseen as part of the Government's new AI initiative, see Section 4.3.

[11] Cf. https://github.com/certainlyio/nordic_bert for Certainly's Nordic BERT models and https://gigaword.dk for a BERT model (called Ælectra) trained on Danish GigaWord.
[12] https://scandeval.github.io
[13] https://www.lr-coordination.eu

Lexical Resources

Lexical and conceptual resources of various kinds are available for Danish, but in some areas resources are still missing. For the general language vocabulary, the largest full form lists and lemma lists are the Orthographic Dictionary,[14] lists based on the Danish Dictionary,[15] and the computational dictionary STO, cf. Braasch and Olsen (2004),[16] all freely available. There are other general wordlists like a lemmatiser dictionary, frequency lists, error spelling lists, synonym lists etc. available from different providers, to be found at sprogteknologi.dk.

With regards to syntactic data for Danish, The Danish Universal Dependencies Treebank (UD-DDT) (Johannsen et al., 2015), which has annotations for dependency structure and part of speech and has recently been annotated with named entities, constitutes a basic resource.[17] The STO lexicon also contains syntactic information such as valency information.[18]

For lexical semantic information on the general language vocabulary, the Danish wordnet, DanNet (Pedersen et al., 2009), is the largest resource with currently around 70,000 concepts and with more than 320,000 semantic relations among them. 8,000 of the concepts are linked to English via Princeton WordNet (Fellbaum, 1998), and the same concepts are linked to multilingual resources, e. g. WordTies[19] and BabelNet.[20] DanNet is currently expanded and will be part of a forthcoming lexical Danish resource, COR (see Section 4.3 for details).

Other semantic resources for Danish are FrameNets which account for the semantic frames and roles related to verbs and deverbal nouns, cf. Nimb et al. (2017) for such a lexicon based on the Berkeley FrameNet standard, and also Bick (2011).

More specific resources are, e. g. three Danish sentiment lexicons, various lists of person names, addresses, place names, and some dialect dictionaries, all available online.[21] For speech there are a couple of publicly available lexicons of transcriptions, the NST Pronounciation lexicon for Danish[22] and the Danpass pronounciation lexicon[23] but most are private.

Regarding Danish terminology, the DANTERMcentre at Copenhagen Business School was for many years a very important contributor to research within methodologies of and approaches to terminology in Denmark. The DANTERMcentre also provided counselling to private and public institutions about terminological principles, and they developed a termbase system which is still in use today in some institutions. In 2016 the DANTERMcentre was partially closed down and counselling and research are no longer performed at the centre.

Terminology resources publicly available today are very few and scattered. As explained below in 4.2.2 Translation Services, an increasing number of public institutions outsource their translation tasks and they do not take the measures to ensure that their terminology is stored, organized and reusable (Kirchmeier et al., 2019). Stakeholders have expressed a strong need for a national termbank, but nothing indicates that such an initiative will be commenced in the foreseeable future. Federated eTranslation TermBank Network (FedTerm)[24] is an EU project that aims to bring together all European term collections in one portal (EuroTermBank). EuroTermBank is thus a potential – although only partial – solution to the non-existing national termbank, but only if Danish public and private institutions decide to participate in this cooperation.

[14] https://dsn.dk/ordboeger/retskrivningsordbogen/ro-elektronisk-og-som-bog/
[15] https://korpus.dsl.dk/resources/index.html
[16] (http://hdl.handle.net/20.500.12115/22
[17] https://github.com/UniversalDependencies/UD_Danish-DDT, Hvingelby et al. (2020)
[18] http://hdl.handle.net/20.500.12115/23
[19] https://wordties.nors.ku.dk
[20] https://babelnet.org
[21] https://sprogteknologi.dk
[22] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-26/
[23] https://schwa.dk/filer/udtaleordbog_danpass/
[24] https://www.european-language-grid.eu/expo-projects/federated-etranslation-termbank-network/

Preprocessing tools

Danish preprocessing tools such as lemmatisers, part of speech taggers, named entity recognisers, and parsers have existed for Danish for several years and are continuously upgraded, partly based on the above mentioned language resources. Some recent achievements for a preprocessing framework made with SpaCy can be found at at Centre for Humanities Computing at Aarhus University, and also Centre for Language Technology has via CLARIN-DK made recent updates on preprocessing of Danish, cf. Jongejan et al. (2021). Further Plank et al. (2020) have made recent advances on data sets for Danish nested named entities (Dan+).

Even if there is still room for improvement, these tools generally achieve high accuracy and are integrated today in most more advanced systems. They are also broadly available through several LT platforms and packages (such as NLTK, SpaCy, CLARIN-DK, sprogteknologi.dk, and Stanford NLP).[25]

4.2 Language Technologies and Tools

Speech Systems and Virtual Assistants

Speech processing includes both speech synthesis (also called text-to-speech systems) and speech recognition (also called speech-to-text systems). Speech recognition has turned out to be a particularly complicated research field when dealing with the tendency of phonetical reduction in modern Danish.

Among Danish companies developing speech recognition technology are for example Dictus ApS and Omilon. They deliver dictation solutions to citizens and many different organizations such as the Danish Parliament, the Danish healthcare system, schools, Danish TV-stations and many more.

Speech technology is also used in chatbots and virtual assistants (or AI assistants as they are also called). The frontrunners of such assistants are Siri (Apple), Alexa (Amazon), Google Assistant (Google) and Cortana (Microsoft) but there are also newer and at least in Denmark, lesser known assistants/bots such as Bixby (Samsung), DataBot (RoboBot Studio), Hound (SoundHound Inc.) and more. Virtual assistants are very popular and perform tasks such as creating text messages, checking reservations, finding hotels, playing music, reading news etc. It is expected that the features, functions, and services of virtual assistants will develop rapidly in the future and that competition will be fierce. Out of all the virtual assistants mentioned above only Siri and Google Assistant work for Danish, and this does not necessarily mean that they work for all variations of Danish.

Danish researchers and other players in this market fear that it is left to foreign tech-companies to decide whether the most popular virtual assistants should be available for Danish. As the Danish language represents only a small market with little or no potential for profit, the necessary technology should instead be developed in Denmark as open source to facilitate the inclusion of Danish in virtual assistants, chatbots and other applications. Currently open-source packages for developing speech recognition for Danish are scarce. An example is DanSpeech (now Alvenir) from DTU (Technical University of Denmark) which is an open-source python package based on the PyTorch deep learning framework.[26] Initiatives like NOTA's development project on speech facilities for people with reading difficulties is also worth mentioning in this context.[27]

[25] https://www.nltk.org, SpaCy https://spacy.io/models/da CLARIN-DK https://clarin.dk/clarindk/tools-texton.jsp, as well as https://sprogteknologi.dk, https://github.com/stanfordnlp/stanza/

[26] https://sprogteknologi.dk/dataset/danspeech

[27] https://sprogteknologi.dk/blog/sprogteknologi-kan-abne-en-verden-af-boger-for-mennesker

Translation Services

Public institutions in Denmark either perform translation tasks within the organization or they outsource the tasks to private translation agencies. Currently most public institutions outsource these tasks, and the tendency is rising (Kirchmeier et al., 2019). No official estimation of the total translation requirements within the public sector is currently at hand, but it is expected that the need for translation services is rising.

Automatic machine translation (MT) is based on several technologies such as big data, neural networks (deep learning), linguistics and cloud computing. Individually – and especially together – these technologies constitute a very high complexity, and there are only few MT systems that include Danish. Two of these systems are Google Translate and Microsoft Translator which have been on the market for years. eTranslation is a newer system that was developed within the framework of the European Commission and offered to the European public sector and SMEs. All three systems are offered as online services – Google Translate and Microsoft Translator are also offered with an api-solution. The translation quality is reasonably high when the translation pair is Danish-English amounting to an average BLEU score of around 0.80 for both Google translate and eTranslation when dealing with domain specific texts, see Nielsen (2022). Translation quality however decreases dramatically when Danish is used in combination with other languages.

Other Language Services

Other language services include a number of specialised technologies such as anonymisation, sentiment analysis, automatic abstracting, summarisation, fake news detectors etc. of which, hardly any currently exist off-the-shelf for Danish. However, services such as opinion mining and sentiment analysis is a growing field since many companies and institutions in Denmark feel an increasing need to monitor and assess opinions and sentiments on the web. For instance, the Alexandra Institute has collaborated with the Danish Broadcasting Corperation on the detection of hate speech.[28] Similarly, a report on the detection of hateful and offensive speech on social media related to politicians in particular was carried out by the company Analyse og Tal in Spring 2021.[29] This report received substantial media attention both for its analysis of social media (which showed that the larger part of Danish social media comments are actually appreciative), but also for being one of the first companies to demonstrate the potential of opinion mining for Danish at a larger scale.

Danish summarisation is also in development with new datasets recently becoming available, as referred to earlier, such as Varab and Schluter (2020).

4.3 Projects, Initiatives, Stakeholders

University Centres, Research and Language Institutes

The Centre for Language Technology[30] was founded in 1991 under the Ministry of Research as the Danish national centre for language technology. To this day, one of its missions (after having merged with the University of Copenhagen (UCPH)) is to carry out and promote strategic research and development in the areas of language technology and computational linguistics in Denmark with particular focus on Danish language resources. Apart from being a university research centre with teaching duties today, the Centre constitutes the Danish

[28] https://github.com/alexandrainst/danlp/blob/master/docs/docs/tasks/hatespeech.md
[29] https://strapi.ogtal.dk/uploads/966f1ebcfa9942d3aef338e9920611f4.pdf
[30] https://cst.ku.dk/english/

Competence Center for LT in the European Language Grid[31] and the Technology National Anchor Point for the European Language Resource Coordination (ELRC). For the last decade, the Centre has hosted the Danish CLARIN (Common Language and Technology Infrastructures) platform, CLARIN-DK[32] for the storage and maintenance of technological language resources in the Danish area.

During the last number of decades, and with a growing need for Danish language technology in the context of AI, quite a lot of additional initiatives have been embarked upon in order to boost Danish LT and NLP further. As a result, dozens of new players have entered the scene both in research, industry and among Danish language institutions.

Several other universities (such as The IT University of Copenhagen, Danmarks Tekniske Universitet (DTU), and Aarhus University among others) and also the computer science department at UCPH have established research centres for NLP and/or language technology, and technological resources are also developed today at the Society for Danish Language and Literature and at The Danish Language Council.

The Alexandra Institute, a private non-profit company that works with research, development and innovation in IT initiated the DaNLP[33] network in 2019 with the aim of supporting the Danish society with more open data resources and tools for Danish NLP.

SMEs working with LT for Danish

As briefly mentioned above, the number of companies in Denmark dealing with Danish LT is increasing quite fast. In fact, recent tentative counts in 2021 indicate that more than 70 companies located in Denmark are working with some degree of substantial development within the area. In some cases, LT constitutes the main focus of a company, in others it is rather a side topic related to a company's interest in adjusting to current technological challenges, related in particular to AI. In the same line of work being carried out at the Alexandra Institute (see above), there is an increasing understanding among these companies that language data is needed, and that the sharing of such data is indeed beneficial. This is also demonstrated in Section 4.2 where a private company was one of the first to release a contextualised so-called BERT model for Danish as open source.

Governmental AI and LT strategies

In 2019, the Danish Government adopted a comprehensive strategy for AI which encompassed an element of Danish LT. In the strategy, ethical and legal issues, more and better data, strong competences and new knowledge, as well as an increased investment in AI and LT became focus areas to which related actions, projects and further initiatives should relate. Following its publication, Denmark has witnessed a surge in activities all related to the development of AI, both in the private and public sector.

When embarking on the strategy, the Danish government combined a focus on both developmental and practical aspects of LT and AI in order to enhance the development and understand the current experience with AI. The previously mentioned report (Kirchmeier et al., 2019, 2020) put together by a language technology committee chaired by the Danish Language Council formed the rationale for the LT component of the investment. Several points of awareness where highlighted in this report:

- The relatively modest investment to date (2019) in research and training in Danish LT in recent years

[31] http://www.european-language-grid.eu/ncc/
[32] https://clarin.dk/clarindk/forside.jspand
[33] https://danlp.alexandra.dk

- The relatively small size of Denmark, both as a language community and as a market and the threat of digital extinction in this context
- The specific characteristics of the Danish language
- The insufficient availability of Danish language resources and datasets
- The insufficient coordination in the development, distribution and use of Danish LT

The recommendations suggested a stronger coordination of efforts as well as the creation of a Danish language bank comprising a series of particular resources. In addition, it was recommended to increase research funding and stimulate the creation of new university study programs in the area of LT and NLP.

To meet these recommendations, the AI strategy now includes an initiative specifically directed towards Danish LT coined "A Common Danish Language Resource". In the 2019 annual joint governmental budget agreement between state, regions and municipals, €2.6 million was invested into the realization of the initiative led by The Danish Agency for Digitisation (for 2019-2026). The first step was the implementation of the aforementioned platform[34] which launched in 2020.

The platform aims to collect metadata and offer easy access to high-quality language resources, counting data, language technology tools and language models. The platform currently displays approximately 130 resources, and enjoyed increased public interest throughout 2021. Furthermore, the agency operating the platform also functions as a point of information for public sector organizations interested in using LT and for actors within the Danish LT community. The agency generally promotes open source development and the sharing of potential (not yet available) language resources.

In addition to collecting and distributing LT resources and tools for Danish, the Agency also creates and funds the development of new LT resources. One of these is constituted by the newly established Central Word Register for Danish, COR.[35] As earlier mentioned, companies and public institutions in Denmark are now working with Danish language data from an NLP and AI perspective. The aforementioned background study indicated an increased request for a standard machine readable lexicon of Danish with basic morphology (lemma, part-of-speech, inflections); semantics (core senses, core ontological typing/supersenses and sentiment information such as positive negative connotation). The COR consortium is formed by three of the core lexical resource players in Denmark: The Danish Language Council, The Danish Society for Language and Literature, and The Centre for Language Technology, University of Copenhagen. By assembling, adjusting and extending existing lexical resources for Danish, the goal is to compile a coordinated lexical resource which meets international standards and where lemmas (including terms) are assigned a unique identifier. Another resource project which will be embarked upon by the Agency in the near future is the development of a large time encoded speech corpus for Danish which will be used for the development of speech technology and integrated in chatbot services etc, cf. the above section for the needs in this field.

Besides the effort to improve the amount of publicly available data, the national strategy is also concerned with the investigation and support with respect to the usage of AI in society. The AI strategy resulted in an investment fund of €25 million (2020 – 2022) which invests in selected AI related projects that seek to improve public services. In 2021, one LT specific project received funding from the investment fund. The project develops speech technology to improve digital citizen support in Aarhus and Roskilde Kommune. In 2022, two further LT specific projects have received funding. In fact, many public and private organizations are

[34] https://www.sprogteknologi.dk
[35] https://sprogteknologi.dk/blog/udarbejdelsen-af-et-centralt-ordregister-skydes-i-gang

already working with AI and LT. For example, forty-eight Danish municipals have invested or are planning to invest in the uptake of chat-bots to improve citizens' services.

Other AI initiatives

Further AI and LT initiatives and actions will come in the near future. Within the 2022 annual state budget, a new digitalization fund received €67 million in funding. These funds are distributed to many different initiatives that support the further digitization of Denmark, which includes uptake of new technologies. Of particular importance in the context of AI, is that €7,3 million have been allocated to the creation of a new data portal that will give researchers and business an overview and access to open data.

Alongside the national investment in Danish LT and AI, several AI research and development initiatives have been launched from mainly private funds to boost AI in Denmark, e. g. approximately €60 million is being invested in 2020 and 2021 through the initiation of two collaborative projects:

- Algoritmer, Data, Demokrati (Algorithms, Data, Democracy)[36] (€13 million funded by The Villum Foundation and the Velux Foundation).
- The AI Pioneer Centre (€44 million) funded by Carlsberg Foundation, Novo Nordisk Foundation, Velux Foundation and Grundforskningsfonden.

Focus in the first project is on the democratic challenges of AI. The project aims to make digital developments work for democratic legitimacy and societal trust. It does so through research, enhanced public understanding of technologies, policy recommendations, and large population surveys etc. The language issue is currently not in focus in the project.

In contrast, the AI Pioneer Centre, which started in late 2021, addresses the development of NLP as a substantial element in AI. The focus is however currently not on language specific (i. e., Danish) issues.

5 Cross-Language Comparison

The LT field[37] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

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[38] broadly categorised into a number of core LT application areas:

[36] https://algorithms.dk

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

[38] 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.

  - 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
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[39]

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

[39] 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.

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[40] 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.[41]

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 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,[42] 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

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

[41] 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.

[42] 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.

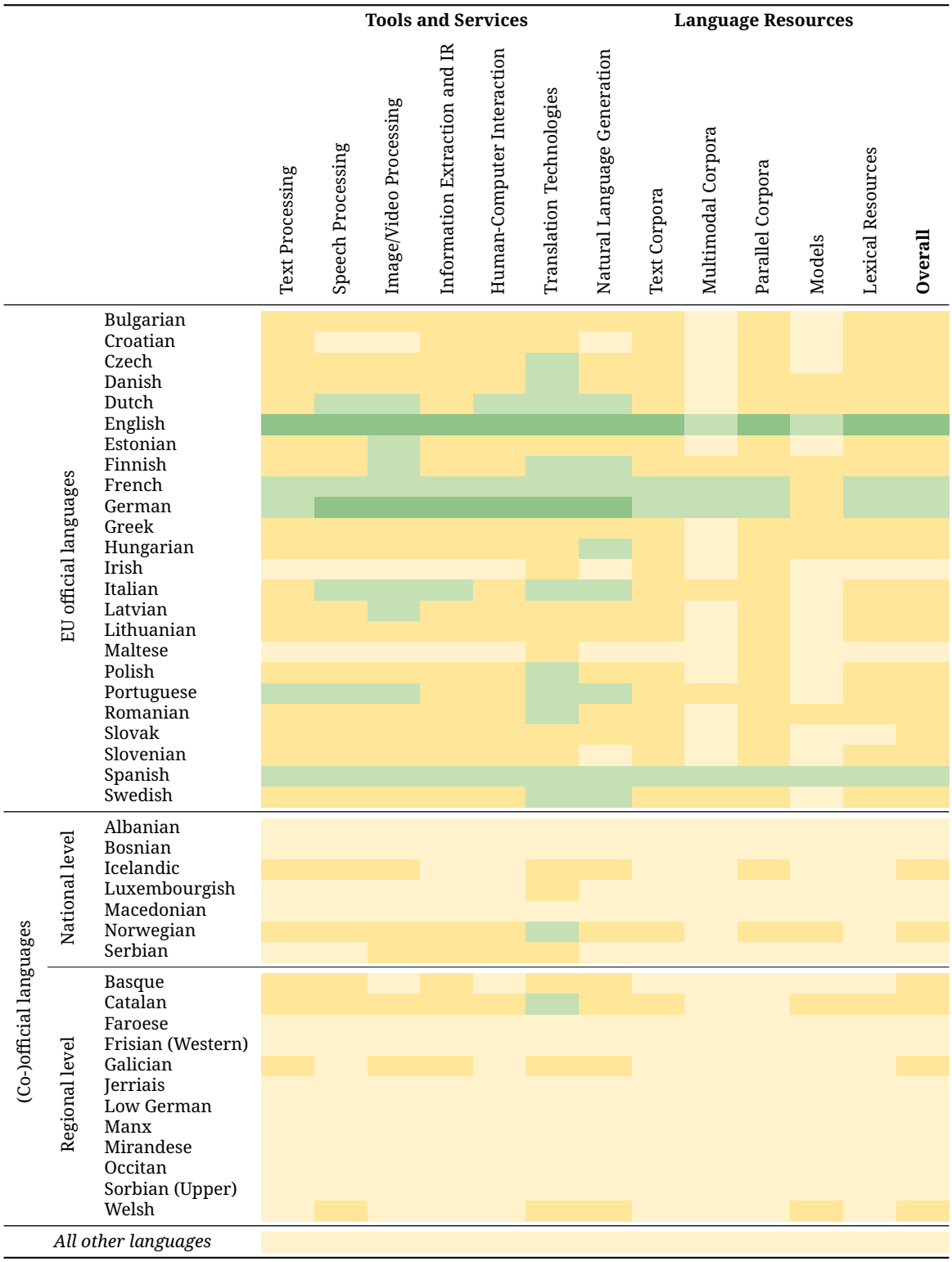

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

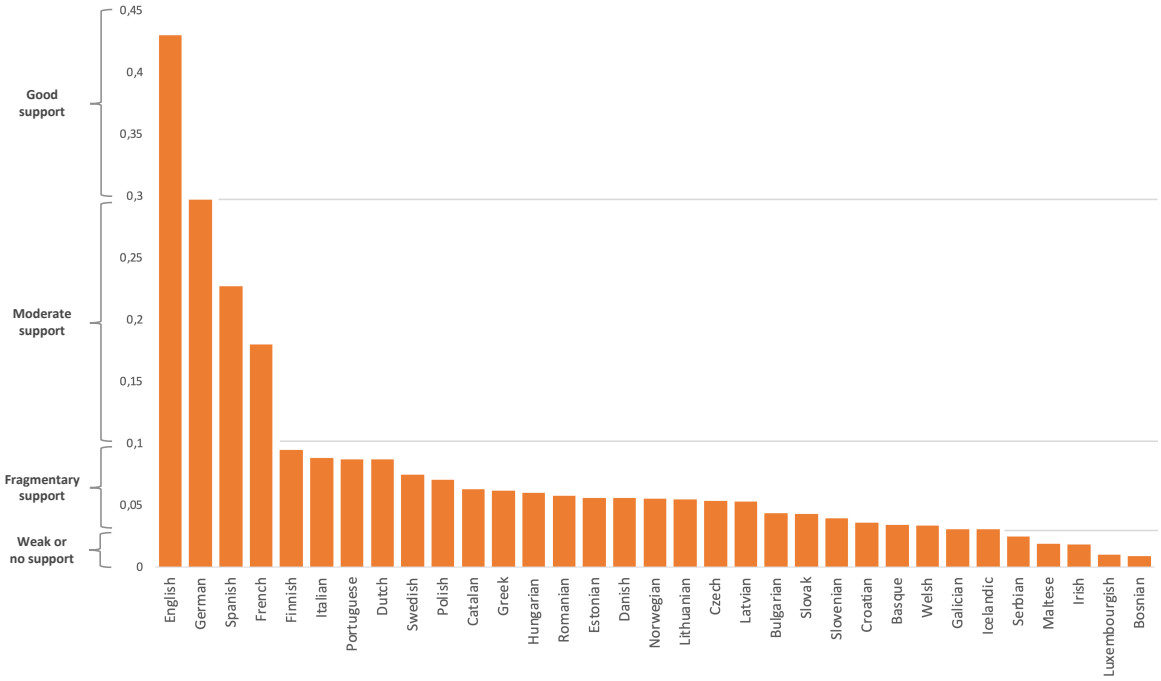

Figure 1: Overall state of technology support for selected European languages (2022)

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 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

Several factors play a role in how fast and how well a language community like the Danish adapts to new technological advances. Even if Denmark is one of the most digitised countries in the world, its relatively small size – both as a language community and as a commercial market – together with our high proficiency of English – seem to have delayed the investments and developments in Danish language processing and LT. The specific characteristics of the Danish language may also play a role, for instance, Danish speech technology is challenged by the tendency of phonetic reduction in spoken Danish, the large number of vowels, by the glottal stop etc.

In this report, however, we have seen a renewed interest in LT at all levels of the Danish society. New stakeholders are emerging day by day together with the increasing tendency of introducing language-centric AI in nearly all aspects of society. With this development comes more focus and better understanding of the challenges of language processing and of why a continuous upgrade of Danish language resources and datasets is indispensable. This increased acknowledgement and tendency of sharing resources across fields is seen both in academia, industry and public administration; it is really good news and will definitely boost LT for Danish in the coming years. As we have seen, new governmental investments are further supporting industry and research in this development. The Danish Agency for Digitisation under the Danish Government is commissioned to host a joint LT platform and to fund and coordinate Danish language resource projects where they are most needed, namely in Danish speech technology and in language understanding, as stated in the Government's AI strategy.

All this being acknowledged, the need for continuous coordinated efforts based in both public institutions, in industrial settings, and in the research community still remains in order to ensure that Danish stays on track to being a digitally fully functional language, also in relation to future language-centric AI solutions.

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.

Eckhard Bick. A FrameNet for Danish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, Riga, Latvia, May 2011. Northern European Association for Language Technology (NEALT). URL https://aclanthology.org/W11-4606.

Anna Rosita Braasch and Sussi Anni Olsen. Sto, A Danish Lexicon Resourse - Ready for Applications. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages

1079–1083. European Language Resources Association, Conference date: 29-11-2010, 2004. ISBN 2-9517408-1-6.

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

Christiane Fellbaum. Towards a representation of idioms in WordNet. In *Usage of WordNet in Natural Language Processing Systems*, 1998.

Henrik Gottlieb. *Echoes of English: Anglicisms in Minor Speech Communities - with Special Focus on Danish and Afrikaans*. Peter Lang, 2020. ISBN 978-3-631-78379-5.

Nina Grønnum. DanPASS - a Danish phonetically annotated spontaneous speech corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1578–1583, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/4_pdf.pdf.

Dorte Haltrup Hansen and Costanza Navarretta. The Danish Parliament Corpus 2009–2017 in CLARIN-DK Centre Repository, 2018. URL http://hdl.handle.net/20.500.12115/8.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. Dane: A named entity resource for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4597–4604, 2020.

Anders Johannsen, Hector Martinez Alonso, and Barbara Plank. Universal Dependencies for Danish. In *Treebank and Linguistic Theories (TLT14)*, pages 157–167, 2015.

Bart Jongejan, Dorte Haltrup Hansen, and Costanza Navarretta. Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus. In M. Monachini and M. Eskevich, editors, *CLARIN Annual Conference*, pages 73–77. Virtual edition, 2021.

Sabine Kirchmeier, Peter Juel Henrichsen, Philip Diderichsen, and Nanna Bøgebjerg Hansen, editors. *Dansk Sprogteknologi i Verdensklasse*. Dansk Sprognævn, 2019. URL https://dsn.dk/wp-content/uploads/2021/01/sprogteknologi-i-verdensklasse.pdf. Contributors from research and industry.

Sabine Kirchmeier, Philip Diderichsen, Peter Juel Henrichsen, Sanni Nimb, and Bolette Sandford Pedersen. World Class Language Technology: The Process of developing a Language Technology Strategy for Danish. *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3290–3294, 2020.

Andreas Kirkedal, Marija Stepanović, and Barbara Plank. FT Speech: Danish Parliament Speech Corpus. *Interspeech 2020*, Oct 2020. doi: 10.21437/interspeech.2020-3164. URL http://dx.doi.org/10.21437/Interspeech.2020-3164.

Jakob Blaaholm Nielsen. Evaluation of Machine Translations from Google Translate and eTranslation: A Quality Assessment of the Machine Translations from English to Danish and vice versa, 2022.

Sanni Nimb, Anna Braasch, Sussi Olsen, Bolette Sandford Pedersen, and Anders Søgaard. From Thesaurus to Framenet. In Iztok Kosem, Carole Tiberius, Milos Jacubicek, Jelena Kallas, Simon Krek, and Vit Baisa, editors, *Electronic Lexicography in the 21st Century*, pages 1–22. Lexical Computing CZ, 2017.

Patrizia Paggio and Costanza Navarretta. The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation*, 51(2):463–494, 2017.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. Danlp: An open-source toolkit for Danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460–466, 2021.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299, 2009.

Bolette Sandford Pedersen, Jürgen Wedekind, Steen Bøhm-Andersen, Peter Juel Henrichsen, Sanne Hoffensetz-Andresen, Sabine Kirchmeier-Andersen, Jens Otto Kjærum, Louise Bie Larsen, Bente Maegaard, Sanni Nimb, Jens-Erik Rasmussen, Peter Revsbech, and Hanne Erdman Thomsen. *Det danske sprog i den digitale tidsalder – The Danish Language in the Digital Age*. 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/danish. Georg Rehm and Hans Uszkoreit (series editors).

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. DaN+: Danish Nested Named Entities and Lexical Normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.583. URL https://aclanthology.org/2020.coling-main.583.

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.

Nina Schneidermann, Rasmus Stig Hvingelby, and Bolette Sandford Pedersen. Towards a Gold Standard for Evaluating Danish Word Embeddings. *Lrec 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 4754–4763, 2020.

Nicolai Hartvig Sørensen and Sanni Nimb. Word2Dict - Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In *Proceedings of the 18th EURALEX International Congres: Lexocography in Global Contexts*, pages 819–827, 2018.

Leon Strømberg-Derczynski, Manuel R Ciosici, Rebekah Baglini, Morten H Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, et al. The Danish Gigaword Corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, 2021.

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.

Daniel Varab and Natalie Schluter. DaNewsroom: A large-scale Danish summarisation dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6731–6739, 2020.