EUROPEAN LANGUAGE EQUALITY

D1.33

Report on the Swedish Language

Authors	Lars Borin, Rickard Domeij, Jens Edlund, Markus Forsberg
Dissemination level	Public
Date	30-04-2022

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	D1.33
Deliverable title	Report on the Swedish Language
Type	Report
Number of pages	27
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 30-04-2022
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	Lars Borin, Rickard Domeij, Jens Edlund, Markus Forsberg
Reviewers	Jane Dunne, 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
ASR	automatic speech recognition
CL	computational linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DLE	digital language equality
ECA	embodied conversational agent(s)
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
ERIC	European research infrastructure consortium
FA	forced alignment
GDPR	General Data Protection Regulation
GPU	graphics processing unit
HPC	high-performance computing
IPR	intellectual property rights
ISOF	Institute of Language and Folklore (*Institutet för språk och folkminnen*)
IVA	intelligent virtual assistant(s)
KB	National Library of Sweden (*Kungliga biblioteket*)
LM	language model
LR	language resource(s)
LT	language technology/technologies
META	Multilingual Europe Technology Alliance
META-NET	EU network of excellence to foster META
ML	machine learning
NLG	natural language generation
NLP	natural language processing
NST	Nordisk Språkteknologi Holding AS
R&D	research and development
SDS	spoken dialogue system
SLT	The Spoken Language Translator
SS	speech science
ST	speech technology
STTI	speech-to-text interpretation
TTS	text-to-speech synthesis
WASP	Wallenberg AI, Autonomous Systems and Software Program

Abstract

Language technology (LT) has a long history in Sweden, with academic research activities going back to the 1950s, undergraduate education in LT since the 1980s, and commercial initiatives taking off in the 1990s. Although there are no dedicated national LT programmes at the moment, there is a large national funding initiative in artificial intelligence (AI), which supports projects that directly benefit LT, such as the building of Swedish deep-learning language models and improved speech technology algorithms. There is a nationally funded research infrastructure with an LT focus, *Nationella språkbanken* (Swedish Language Bank), which also administers Swedish membership in CLARIN ERIC, as well as several significant national research projects and new organisations either dedicated to or relevant for LT.

Swedish is relatively well-endowed with language resources and language tools, but there are also numerous gaps that need to be filled. In particular, we must come to terms with the methodological sea change brought to LT by the recent rise to prominence of deep learning LT systems (often under the guise of AI). The so-called *language models* coming out of such systems present black-box solutions which achieve state of the art performance on several LT problems – in particular in natural language understanding – despite being trained on raw, unlabelled language data. In the Swedish context, some computer science centres have started showing an interest in LT as a central component of AI, more often than not without awareness of the long history and significant accomplishments of Swedish LT. In addition, both *commercial enterprises* and *public institutions*, other than universities, are showing an interest in developing language-aware applications for Swedish.

The Swedish academic LT expertise represents seventy years of accumulated knowledge, and is characterised by a well-balanced mix of researchers from computer science and linguistics (engineering and phonetics in the case of speech technology), which we believe constitutes a valuable knowledge base which should not be allowed to erode. Short-term, the best way of ensuring this is probably to focus on language resource development, where much work still remains to be done for Swedish. Well-designed gold-standard corpora for fine-tuning language models and evaluating LT systems will require exactly this kind of expertise for their construction, not least in order to avoid pitfalls such as models making undesirable biased predictions that risk perpetuating gender roles or lead to unfair treatment of minority groups.

In the medium term, we should aspire to understand the internal workings of current language models better (in the spirit of the emerging research field “explainable AI”), in order to be able to exploit already existing linguistic knowledge (for instance, information about words collected in a lexical or conceptual resource) when training language models. This will potentially reduce their training data requirements, thus putting state of the art LT tools in reach of lower-resourced languages (including the official minority languages and widely-spoken immigrant languages of Sweden).

It is clear that several decades of focused work on achieving funding for a collected push for Swedish language and speech technology resources, together with the upswing of new deep-learning methods for the same technologies, has paid off. Consequently there is currently a national research infrastructure for LT, as well as several significant research programmes, national projects and new organisations either dedicated to or relevant for LT.

At the same time, recruiting highly skilled LT engineers, developers and researchers is difficult. In the future, we would like to see even closer collaborations and communication between the “traditional” LT research community and the new AI field, e.g., through the establishment of dedicated academic LT training programmes on all levels and from earmarked national funding for LT research.

Sammanfattning

Språkteknologi är ett samlingsnamn för sådan informations- och kommunikationsteknologi som låter datorer hantera mänskligt språk i alla dess former – tal, skrift och teckenspråk. Det är ett starkt tvärvetenskapligt forskningsområde som är relevant överallt där människor interagerar med datorer och även vid interaktion människor emellan, i form av olika sorters kommunikationshjälpmedel.

Den svenska forskningen i språkteknologi har en lång historia. På våra universitet har det funnits som forskningsområde sedan 1950-talet, specialiserad grundutbildning i språkteknologi startade på 1980-talet och kommersiella initiativ tog fart under 1990-talet. Den nationella forskarskolan i språkteknologi (GSLT: Swedish National Graduate School of Language Technology), som verkade med nationell finansiering under 2000-talets första decennium och producerade omkring 50 doktorer med språkteknologisk inriktning i olika akademiska discipliner, har haft stor betydelse för att skapa en nationell intressegemenskap bland forskarna i ämnet.

Det offentliga stödet för forskning och utbildning i språkteknologi har varierat över tid. För tillfället finns inga nationella forskningsprogram med den inriktningen, men eftersom språkteknologi är en av grundstenarna i *artificiell intelligens* (AI), är det mycket positivt att det finns ett stort Wallenbergfinansierat forskningsprogram inom AI, WASP, som stödjer projekt som direkt gynnar språkteknologi, såsom utveckling av svenska så kallade djupinlärande språkmodeller och av förbättrad talteknologi. Vidare finns en nationellt finansierad forskningsinfrastruktur med språkteknologifokus, *Nationella språkbanken*, som också administrerar det svenska medlemskapet i den europeiska forskningsinfrastrukturen CLARIN ERIC. Dessutom har svenska språkteknologiforskare i fri konkurrens kunnat säkra ett antal betydande nationella forskningsprojekt med språkteknologiskt fokus eller på annat sätt relevanta för området (finansierade av bl.a. Vetenskapsrådet, Vinnova och Riksbankens Jubileumsfond).

Språkteknologi har både språkoberoende och språkberoende aspekter. Detta betyder att resultat som kommer ur språkteknologisk forskning om svenska är högst relevanta för den internationella forskargemenskapen, men också att språkteknologi för svenska inte kommer till utan vidare; den måste skapas i Sverige. Språkteknologi har vittgående betydelse för svenskans framtid som fullödigt språk. Informationssamhället avancerar på bred front, och utan språkteknologi för ett språk kan man inte räkna med att upprätthålla önskvärd tillgång till digital information eller digitala tjänster på det språket. Här blir även flerspråkiga lösningar viktiga eftersom man vill kunna hantera så många som möjligt av Sveriges språk och även hantera det faktum att var femte invånare i Sverige är född någon annanstans.

För att utveckla språkteknologi för ett språk krävs både så kallade språkresurser (textsamlingar, taldatabaser, digitala lexikon, etc.) och språkverktyg för olika former av analys och bearbetning av språk. Tack vare Sveriges långa historia av nationellt samordnad språkteknologiforskning är svenskan relativt välutrustad med språkresurser och språkverktyg, men det finns också många luckor som måste fyllas. Viktigt i det sammanhanget är att den nya ”AI-revolution” som vi möter dagligen i media bland annat har inneburit en grundläggande förändring av hur språkverktyg utvecklas. Tidigare handlade det huvudsakligen om att formulera och programmera formella regler för språkanalys, medan dagens AI-system bygger på *maskininlärning*: de lär sig de relevanta regelmässigheterna om de förses med rätt sorts träningsdata. Många språkförståelseproblem hanteras idag i praktiken bäst med så kallade djupinlärande system som matas med enorma kvantiteter (miljarder ord) ren text eller tal. Språkmodellerna som kommer ut ur sådana system är dock i princip komplexa svarta lådor: det är inte känt hur deras interna tillstånd hänger ihop med hur språkvetare brukar beskriva språket.

I Sverige har uppsvinget för AI även lett till att vissa datavetenskapliga forskningsmiljöer

börjat visa intresse för språkteknologi, men ofta utan medvetenhet om forskningsområdets långa historia och signifikanta landvinningar i Sverige. Dessutom visar både kommersiella företag och andra offentliga aktörer än universitet intresse för att utveckla språkteknologiska tillämpningar på svenska för sina specifika behov.

Den svenska akademiska expertisen inom språkteknologi representerar sjuttio år av mödosamt ackumulerad kunskap och kännetecknas av en väl avvägd blandning av forskare från datavetenskap och språkvetenskap (ingenjörsvetenskap och fonetik när det gäller talteknologi), som utgör en omistlig kunskapsbas som bör inte tillåtas vittra sönder. På kort sikt är det bästa sättet att säkerställa detta förmodligen att fokusera på språkresursutveckling, där mycket arbete fortfarande återstår för svenskan. Väl utformade referenskorpusar för att trimma in språkmodeller och utvärdera språkteknologisystem kommer att kräva just denna typ av expertis för sitt uppbyggande, inte minst för att undvika fallgropar, till exempel i form av skevhet i språkmodellerna som riskerar att vidmakthålla könsroller eller leda till orättvis behandling av minoritetsgrupper.

På medellång sikt bör vi sträva efter att förstå de nuvarande språkmodellernas interna funktion bättre (i anslutning till det framväxande forskningsområdet ”förklarbar AI”), inte minst för att kunna utnyttja redan existerande språkkunskaper och högvärdiga språkresurser (till exempel information om ords formella beteende och deras semantik som samlats in i en lexikonresurs) när man tränar språkmodeller. Detta kan minska deras krav på träningsdata, vilket gör att de senaste språkverktygen kan bli tillgängliga även för språk med färre resurser (inklusive de officiella minoritetsspråken och de stora invandrarspråken i Sverige).

Vi kan konstatera att flera decennier av fokuserat arbete med att skaffa finansiering för en samlad satsning på svenska språk- och talteknologiska resurser, tillsammans med introduktionen av nya djupinlärningsmetoder för dessa teknologier, har gett resultat, och följaktligen finns det idag en nationell forskningsinfrastruktur, samt ett antal nationella forskningsprojekt för språkteknologi. Samtidigt är det svårt att rekrytera utbildade forskningsingenjörer, utvecklare och forskare inom området. I framtiden skulle vi vilja se ännu närmare samarbeten och kommunikation mellan det ”traditionella” språkteknologiforskarsamhället och det nya AI-området, t.ex. genom etablering av specialiserade utbildningar på alla nivåer och genom öronmärkta nationella medel för språkteknologiforskning.

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.[1]

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

[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.

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 Swedish Language in the Digital Age

Swedish is the main language of Sweden. In addition, since the year 2000, five languages are officially recognised as national minority languages in Sweden: Finnish, Yiddish, Meänkieli, Romani and Sami.[2] The status and rights for the languages in Sweden are stated in the *Language Act*,[3] which also officially recognises Swedish Sign Language, and the *Act on National Minorities and National Minority Languages*.[4] Swedish is also the official language of Åland (an autonomous region of Finland), the second constitutional official language of Finland,[5] and since 1995, an official language of the European Union. Small pockets of Swedish speakers in Estonia and Ukraine are all but extinct, although traces can still be found. Conversely, a recent diaspora has emerged in Norway, with estimates of the number of Swedish residents ranging from 10,000 (Wessel et al., 2018) to 50,000.[6]

No official statistics are kept regarding the languages of residents of Sweden (Parkvall, 2019), and numbers have to be estimated from other sources. There are about 10 million native speakers of Swedish, the vast majority of which are Swedish citizens,[7] and *Ethnologue* (Eberhard et al., 2021) lists another 3.2 million second-language speakers.[8] According to Statistics Sweden, in 2021 2,090,503 residents, or 20% of the Swedish population, were born outside Sweden, and for another 6.3% both parents were born outside Sweden.[9]

Swedish is spoken in all levels of government and education in Sweden and on Åland. Its vitality is strengthened by its closeness to the languages spoken in neighbouring Norway and Denmark: speakers of Swedish, Norwegian and Danish are able to communicate with relative ease (Haugen and Borin, 2018). Together, these languages have around 20 million native speakers.

Swedish is written using a modified Latin script with a 29-letter alphabet (the 26-letter basic Latin alphabet is extended with the vowel characters <å>, <ä> and <ö>).[10] The writing system is in the mid-range of orthographic transparency.

In general, Swedish is a relatively normal representative of European languages, and Germanic languages in particular. The most “exotic” aspects of the language are found in the domain of phonology, with notable features being: a phonemic pitch accent system; the cross-linguistically rare phoneme /ɧ/; an unusually large vowel system, including front rounded vowels (where the high vowels display a notable two degrees of rounding: /ʉ̟ y/); and rather liberal phonotactics with CCC onsets and CCCC codas, yielding half a million potential syllables. Structurally, Swedish generally follows the patterns typical of Germanic languages, including V2 word order. Among more unusual traits we find negation placement before the tensed verb in subordinate clauses, a “reflexive possessive” in the third person (i.e., a spe-

[2] See further the ELE report on Nordic minority languages (Moshagen et al., 2022).
[3] SFS 2009:600, https://rkrattsbaser.gov.se/sfst?bet=2009:600
[4] SFS 2009:724, https://rkrattsbaser.gov.se/sfst?bet=2009:724
[5] Around 5% of the Finnish population (roughly 300,000 including 30,000 on Åland) have Swedish as their mother tongue; see https://www.tilastokeskus.fi/tup/suoluk/suoluk_vaesto_en.html
[6] https://sverige-norge.se/var-tionde-oslobo-ar-nu-svensk/
[7] https://www.isof.se/lar-dig-mer/kunskapsbanker/lar-dig-mer-om-svenska-spraket/om-svenska-spraket
[8] Swedish is taught at a large number of schools and universities over the world, and the Swedish government authority, the Swedish Institute, collaborates directly with around 200 teaching institutions in about 40 countries worldwide: https://svenskaspraket.si.se/sa-arbetar-vi/universitet-med-svenskstudier/.
[9] https://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/befolkningens-sammansattning/befolkningsstatistik/pong/tabell-och-diagram/helarsstatistik--riket/befolkningsstatistik-i-sammandrag/
[10] Note that these additions are treated as distinct letters, and not as variants of <a> and <o>, as testified by their placement at the end of the Swedish alphabet, after <z>.

cial possessive form used if and only if the possessor is co-referential with the subject), and the recent introduction (and wide adoption) of a consciously coined gender-neutral third-person singular personal pronoun (*hen* 'he/she'), resulting in a five-member set of personal pronouns in the third person singular:

HUMAN			NON-HUMAN	
han 'he'	*hon* 'she'	*hen* 'he/she'	*den* 'it' (NON-NEUTER)	*det* 'it' (NEUTER)

Dialects and minority languages

Parkvall (2009) estimates about 185,000 native speakers of highly divergent Swedish dialects, of whom 5–10,000 use varieties divergent enough from the standard language to merit being considered (indigenous minority) languages in their own right.[11] In general, however, the regional differences in Sweden are moderately marked, and – as in most other industrialised countries – people born after the Second World War generally speak the standard. Differences betraying approximate geographical origin mainly concern the phonology, phonetics and prosody, with few lexical peculiarities. Swedish-speakers in Finland have generally followed the same path, although the local dialects are somewhat more distinct than they are in Sweden. However, east of the Bay of Bothnia, words and constructions denoting concepts regarding modern society are frequently borrowed or calqued from Finnish. The geographical differences that do exist are in effect exclusive to the spoken language, and for a newspaper text, it would be virtually impossible to determine the area in which it was produced, and even for a newspaper from Finland, this would be difficult, save for a small number of words and expressions denoting concepts relating specifically to Finnish society.

Swedish in the digital sphere

Sweden belongs to the group of European countries in which 95% or more of the population use the internet at least once a week.[12] In 2020, 86% of Swedish households were connected to 100 Mb or faster fibre optic and 93% of populated areas had stable and fast cell phone coverage (Ingman, 2021), and 90% of the population used a smartphone.[13] 15% of households had at least one smart speaker in 2018, while the corresponding number in the US was 21%. The poorer quality of the Swedish speech technology as compared to the English is pointed out as part of the explanation.[14] 82% of Swedes were on social media in 2021.[15]

Over the last 5 years, the .se top domain has had, at any given point in time, somewhere between 1.5 and 2 million registered domain names. The top domain .nu is also a popular choice for Swedish sites.[16] The .nu top domain had between 250,000 and 500,000 registered domain names in the same time period.[17] Swedish web pages are overwhelmingly produced in Swedish, and quite often an English translation is provided, at least for parts of the material. This is in line with the general situation: the majority of mainstream software such as operating systems, word processors, etc., is localised to Swedish, although poor translations

[11] For information on Sweden's official minority languages, see the ELE report on Nordic minority languages (Moshagen et al., 2022).
[12] https://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do; the others were, in 2021, the other four Nordic countries plus Ireland, Switzerland and Luxembourg.
[13] https://www.statista.com/statistics/568272/predicted-smartphone-user-penetration-rate-in-sweden/
[14] https://www.telecompaper.com/news/nearly-one-sixth-of-swedish-homes-have-a-smart-speaker-google-has-55-share--1279546
[15] https://datareportal.com/reports/digital-2021-sweden
[16] Due to a Swedish version of domain hacking: *nu* means 'now' in Swedish.
[17] https://internetstiftelsen.se/domaner/domannamnsbranschen/domanstatistik/

are still a recurring source of irritation and mirth. Some international online shopping platforms are also localised to Swedish, often using unsupervised machine translation, which leads to the occasional outbreak in social media.[18]

3 What is Language Technology?

Language[19] is the most common and versatile way for humans to convey information. We use it 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) or *natural language processing* (NLP). There are differences in focus and orientation, since CL tends to be more informed by linguistics and NLP by computer science.

Alongside CL and NLP, both of which have historically and contemporarily focused on written language, the highly interdisciplinary field of *speech technology* (ST) aims to capture, study, analyze, model, synthesise and generate spoken language and spoken interaction. Although there are obvious connections and similarities between CL/NLP and ST, they take place largely in different research communities, at different institutions and organisations; they are published at different venues; and they use different resources and technologies.[20] ST should not be overlooked, however, as many of the poster technologies of language-centric AI, such as smart assistants, social robots, tele-presence, collaborative computers and manufacturing robots, health assistants and trackers of for example dementia, and language tutors, are in fact examples of speech-centric AI. Alongside ST we find *speech science* (SS). The relationship between the two resembles that between NLP and CL, in that ST is more informed by computer science and SS by linguistics and phonetics.

Language technology (LT) is used here as a more neutral term, covering CL, NLP, ST and SS. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably AI), mathematics and psychology among others. When spoken or signed human languages are concerned, the list includes a variety of additional fields, such as phonetics, phonology, signal processing, acoustics, physics, physiology, mechanics, engineering, anatomy, aerodynamics, robotics, planning, interaction analyses, and computer vision.

LT 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 signed.

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 language 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 analyzed and/or produced. Gradually, with the evolution and

[18] A particularly well published example is the translation of *rape* in food products containing rape seed to *våldtäkt*, the 'sexual assault' sense of *rape*.

[19] This section presents a slight revision of a text provided by the editors, which in turn is an adapted summary of the ELE Deliverable D1.2 Report on the state of the art in language technology and language-centric AI (https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf) and of sections 1 and 2 of the ELE Deliverable D3.1 Report on existing strategic documents and projects in LT/AI (https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revise_.pdf)

[20] A main difference is that ST regularly involves studies of humans and their actions, as speech and spoken communication is a largely interactive and emergent phenomenon. For the same reason, it also commonly involves several modalities such as facial expressions, head movements, and gestures.

advances in machine learning (ML), rule-based systems have been displaced by data-based ones, i.e., systems that learn implicitly from examples. In the recent decade of the 2010s, we have observed a radical technological change in both NLP and ST: 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 text words (called *word embeddings*) using vast amounts of unlabelled data and using only some labelled data for fine-tuning. We are now 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 LT processing large and diverse multilingual data); (3) increase in high performance computing (HPC) power in the form of graphics processing units (GPUs); and (4) application of simple but effective self-learning approaches.

LT strives to provide solutions for the following main application areas:

- **Text analysis** which aims at identifying and labelling the linguistic information underlying any text in written in human language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of word morphology, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at analyzing, understanding and generating unimodal or multimodal speech. Some of the main technologies are speech synthesis, that is the generation of speech, given either a piece of text (i.e. text-to-speech synthesis, TTS) or some other stimuli such as an intent or a situation, and automatic speech recognition[21] (ASR), that is the conversion of a captured speech signal into a written transcription, forced alignment (FA), the alignment in time of speech and its transcription, attitude and emotion recognition, and speaker recognition and verification, which aim at finding out a speaker's identity.
- **Spoken interaction processing** aims at allowing humans to communicate with each other and with electronic devices through spoken language through the modelling, understanding and generation of (contributions to) face-to-face or mediated spoken interaction. Technologies include spoken dialogue systems (SDS), embodied conversational agents (ECA), intelligent virtual assistants (IVA) and social robotics.
- **Machine translation**, the automatic translation from one language into another. The standard application is translation of writing, and special cases include (realtime) speech-to-speech translation, (realtime) speech-to-text interpretation (STTI), and (realtime) translations between, to and from sign languages.
- **Information extraction and retrieval** which aim at extracting formally structured information from text documents, finding appropriate pieces of information in large collections of text material, such as the internet, and providing documents, text snippets, or videos and audio recordings of speech that include the answer to a user's query.
- **Natural language generation (NLG)**. The task of automatically generating written texts. Summarisation, such as the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.

[21] ASR is the oldest and most broadly used term in the ST community. Other terms include speech-to-text (conversion) and automatic transcription.

LT already imbues our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just 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.[22] These reports are unclear as to the extent to which they include LT, but the intelligent virtual assistant market alone was reported at a global market value of USD 5 billion in 2020, with a projected annual growth rate of 30% for eight years, reaching USD 50.9 billion in 2028 (https://tinyurl.com/2p9c3xe3).

4 Language Technology for Swedish

4.1 Availability of Language Data and Tools

Text corpora

Monolingual text corpora There is a wealth of monolingual text corpora available for Swedish.[23] Korp, the corpus infrastructure of Språkbanken Text,[24] at the time of writing provides online access to 274 monolingual corpora of Present-Day (20th and 21st century) Swedish, with a total of about 14.5 billion tokens.[25] Most of the corpora have been automatically linguistically annotated with sentence boundaries, with lemmas and morphosyntactic descriptions of words and multi-word expressions, named-entity information, dependency syntax, and word senses from a large Swedish semantic lexicon. The corpora can also generally be downloaded in a simple XML format containing full linguistic annotations from Språkbanken Text's resource pages,[26] although for intellectual property rights (IPR) reasons many of them are sentence-shuffled.

Reference corpora There are not so many gold-standard text corpora available for Swedish, however. Those that do exist were mostly compiled some time ago and thus do not completely reflect present-day vocabulary and – more importantly – also lack web material, e.g. the *Talbanken* dependency treebank (Berdicevskis, 2020) (95k tokens) with non-fiction texts

[22] In a recent report from 2021, the global NLP market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://tinyurl.com/2p9ed6tp). The report defines *NLP* as "a part of computer science and artificial intelligence that deals with computer-human language interaction". 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).

[23] Note that the term *corpus* as used in LT is considerably wider than its use as a technical term in corpus linguistics, so that most of the datasets described in this section would not qualify as corpora in the latter field. In this report, we will stick to the usage adopted in LT, basically that of 'processed collection of text(s)'.

[24] https://spraakbanken.gu.se/korp

[25] Språkbanken Text is a CLARIN B-centre. Hence, it provides access to Swedish text corpora produced by several groups from across Sweden and from Finland.

[26] https://spraakbanken.gu.se/en/resources/corpus

from the 1970s or the Stockholm Umeå Corpus (1.2M tokens) containing a Brown Corpus-style genre-balanced selection of texts from the early 1990s with various word-level annotations (lemmas, morphosyntactic descriptions and named-entity annotations). Notable exceptions are, e.g., the Eukalyptus treebank[27] (100k tokens) compiled by Språkbanken Text, and a gold-standard corpus for Swedish named-entity recognition under development under the auspices of Swe-Clarin (Ahrenberg et al., 2020), both containing only freely available texts. Finally, there is currently an ongoing national collaboration with the aim of creating a Swedish natural language understanding benchmark like (Super)GLUE (Wang et al., 2018, 2019),[28] called SuperLim[29] (Adesam et al., 2020). Version 2.0 of SuperLim will be released at the end of 2022.

Bi- and multi-lingual text corpora Språkbanken Text offers access to parallel text corpora in Swedish aligned with 27 other languages.[30]

A collection of parallel texts collected from Swedish public agencies are made freely available via Nationella språkbanken[31] and ELRC SHARE,[32] in the form of either (1) translation memories, i.e. sentence-aligned texts in different languages that have either been extracted from translation tools or from automatically sentence-aligned parallel texts; or (2) parallel texts in several languages, that either have been crawled from public agency web sites or that were received directly from the agency. The repository contains approximately 1,700 texts in 40 languages. The original Swedish texts contain a total of 1.4 million tokens. Most of these texts have been translated into English, and different subsets of the texts have been translated into subsets of the other 38 languages.[33]

Spoken-language and speech corpora

Written speech (orthographic or phonetic transcriptions of speech) There are few publicly available collections of substantial amounts of transcribed speech. The following two resources deserve mentioning: (1) **The transcriptions of the Swedish parliament**: like many other countries, Sweden makes both recordings and transcripts of parliamentary discussions and debates freely available, in several ways;[34] (2) **Swedish subtitles**: at the national broadcaster – Swedish Television, a very large proportion of its Swedish-language programming is subtitled (80%).[35]

In both these cases, the texts are not literal transcriptions, but rather interpretations that follow guidelines and practices that have often not been formalised or documented, and that vary over time (e.g. Ljusterdal, 1973; Hallberg, 1994).

Spoken text (read speech) The earliest monolingual Swedish speech corpus to be recorded specifically with speech technology applications in mind is the EU-funded Swedish SpeechDat which was recorded in the Spoken Language Translator (SLT; Rayner et al., 2000) project. The corpus has not been made publicly available as release forms are missing or fail to meet current standards.

[27] https://spraakbanken.gu.se/en/resources/eukalyptus
[28] https://gluebenchmark.com, https://super.gluebenchmark.com
[29] https://spraakbanken.gu.se/en/resources/superlim
[30] https://spraakbanken.gu.se/korp/?mode=parallel, https://spraakbanken.gu.se/en/resources/corpus?s=parallel&language=All
[31] https://www.sprakbanken.se/sprakbankeninenglish.html
[32] https://elrc-share.eu
[33] https://snd.gu.se/en/catalogue/collection/parallel-texts-from-public-agencies. See Skeppstedt et al. (2020).
[34] See, e.g., https://data.riksdagen.se/in-english/ and Rauh and Schwalbach (2020)
[35] See https://kontakt.svt.se/guide/undertext

Shortly after, around the turn of the century, Norwegian speech technology company Nordisk Språkteknologi Holding AS (NST) recorded several sizeable Swedish speech corpora for speech technology, where an attempt was made to cover regional variation through balancing over 10 dialect areas. The company went bankrupt in 2003, and its speech resources were eventually made publicly available under a CC-ZERO license through the Norwegian Language Bank (Andersen, 2005, 2011; Norwegian Language Bank, 2020; Mossberg, 2016; Vanhainen and Salvi, 2014).

In recent years, the need for Swedish speech data has again become apparent, and a number of projects now aim to record and make available speech corpora. Notably, Språkbanken Tal is creating an ASR corpus with 100 speakers recorded in a studio setting, as well as recordings for a male and a female TTS voice, and the Finnish national broadcaster Yle, in collaboration with the Finnish Language Bank, are recording Finnish Swedish voices donated by the public in the project *Donera prat* 'Donate speech'. Both these datasets will be available under an open license.

Apart from the SLT and NST recordings, which contain a measure of dialectal variation, there are no specifically speech technology oriented corpora focusing on regional variation, but a few other speech collections with a dialect focus exist.

Spoken speech (unscripted speech) A wide range of ST tasks aim to analyse, understand, model, or generate speech as it occurs in real-world interactions. Perhaps more importantly, a wide range of speech-centric sciences aim to understand and explain the inner workings of human spoken interaction. Large annotated corpora of representative, unscripted, real-world speech would make up the ideal foundation for both of these areas, but in practice, these do not exist. For Swedish, the lack of freely available recordings of real-world speech is an inhibiting factor for speech technology development beyond relatively simple and controlled applications and domains. At the same time, the de facto availability of unannotated audio and video recordings of speech on the internet is greater than ever. This type of data is sometimes referred to as *found data* (roughly 'data that was acquired for some other reason than for which it is now used'). Unfortunately, the legality and circumstances under which the use of found data is permissible are especially unclear when speech is involved.

Multimodal and sign-language corpora

There is a distinct lack of publicly available large multimodal corpora specifically designed for or curated for speech- and/or language technology purposes. Still, there are several collections that can be, and are, used for such purposes. Among the largest and richest available are **parliamentary data**, which deserve mention as a multimodal resource as well, as they combine a large number of speech and video recordings with transcripts with a large number of other parliamentary documents; and **the audiovisual collections at the National Library of Sweden (KB)**, by far the largest audiovisual collection in Sweden, and continuously growing.

The Sign Language Research Unit at Stockholm University provides access to a **Swedish Sign Language corpus**, with close to 200k annotated tokens.[36]

Lexical and conceptual resources

Language resource compilation and LT for written Swedish got started in the 1960s largely motivated by lexicographic considerations. Ever since, work on lexical resources for Swedish has been informed by and has informed Swedish LT in a virtuous circle. For this reason,

[36] https://www.ling.su.se/teckensprakskorpus, http://sts-korpus.su.se

Swedish is well-equipped with high-quality lexical and conceptual resources. Språkbanken Text maintains *Swedish FrameNet++* (SweFN++; Dannélls et al., 2021)[37], a large lexical macro-resource which interlinks a number of lexical and conceptual resources, e.g., SALDO (a conceptual resource with almost 150,000 word senses), Swedish FrameNet (its approximately 42,000 lexical units make it the largest framenet in the world in this respect), a Swedish sentiment lexicon, several multilingual lexicons, and a Swedish version of Roget's *Thesaurus*. A notable lacuna in SweFN++ is a Swedish wordnet, which is still pending.

ISOF publishes Lexin, a dictionary series from Swedish to more than 20 immigrant languages with focus on words for the Swedish society (5,000 or 28,000 lemmas depending on language). Digital Lexin is currently available in the following 18 languages: Albanian, Amharic, Arabic, Azerbaijani, Bosnian, Finnish, Greek, Croatian, Northern Kurdish, Pashto, Persian, Russian, Serbian, Somali, Spanish, Southern Kurdish, Tigrinya and Turkish. Data are freely available via Nationella språkbanken.[38]

ISOF publishes term collections in the National Term Bank of which some are freely available via Nationella språkbanken.

Models and grammars

For **text processing**, grammar-based LT has largely yielded ground to machine-learning approaches based on deep learning neural approaches.[39] The National Library of Sweden (KB) have taken a leading role in training Swedish language models (LMs), which makes good sense since it is an organisation which by law has access to everything published in Sweden.[40] There is a plethora of different kinds of LMs, but the most popular one is BERT (Devlin et al., 2019). The Swedish one, trained by KB, has 54k downloads/month at the time of writing. This can be compared with the English BERT with 15.7m downloads/month.

For **speech processing**, several Swedish acoustic models for Kaldi are available for Swedish. A number of wav2vec models have also been trained recently,[41] and freely available models based on controlled datasets are being developed by Språkbanken Tal.

Tools and services

Tools and toolchains for **text processing** are being offered by several academic centres in Sweden, e.g., Språkbanken Text's Sparv corpus annotation pipeline,[42] Robert Östling's (Stockholm University) NLP tools,[43], and SWEGRAM,[44] developed by the language technology research group at Uppsala University. The component tools of such toolchains are at present almost without exception based on some form of machine learning, typically fine-tuning of a general neural LM.

Signal processing tools for **speech** are largely language independent. Notable Swedish tools include Wavesurfer[45] and the Snack Sound Toolkit[46]. Swedish language support is also

[37] https://spraakbanken.gu.se/en/research/themes/swedish-framenet-plus-plus. SweFN++ can be downloaded under a CC-BY license from Språkbanken Text's resource pages: https://spraakbanken.gu.se/en/resources/lexicon
[38] https://snd.gu.se/en/catalogue/study/ext0286
[39] Although in areas such as controlled languages grammar-based LT is still going strong, for example: https://www.grammaticalframework.org
[40] https://huggingface.co/KBLab
[41] See for example https://huggingface.co/KBLab
[42] https://spraakbanken.gu.se/sparv
[43] https://github.com/robertostling
[44] https://cl.lingfil.uu.se/swegram/en/
[45] https://sourceforge.net/projects/wavesurfer/
[46] Included in a number of Linux distributions, see https://en.wikipedia.org/wiki/Snack_Sound_Toolkit

included in multilanguage toolkits such as the Montreal Forced Aligner[47] and the BAS Web Services[48].

There is a growing interest in combinations of **image/video processing** and LT. Examples include a WASP focus area to link entities of sentences with objects in a different modality[49] and efforts to link subtitles with video.

Especially notable is the work at Stockholm University on developing LT tools for (transcribed) **Swedish Sign Language** (Östling et al., 2015; Östling et al., 2017).

Research on **machine translation and computer-aided translation** involving Swedish is conducted at the University of Helsinki in Finland,[50] and at Uppsala University. Commercial translation tools and services handle translation between Swedish and English reasonably well, and the Helsinki group has worked extensively on the language pair Swedish–Finnish,[51] but there is a definite need for translation tools dealing with Swedish as source language and the other national minority languages and the largest immigrant languages as target languages.[52]

Research on **(spoken) dialogue systems**, including spoken interaction with robots, is pursued at least at the University of Gothenburg and at KTH Royal University of Technology (Stockholm), and Swedish researchers in the field have been involved in the build-up of companies in the field, several of which have internationally leading positions, such as Voice Provider, Artificial Solutions, Talkamatic, and Furhat Robotics.

Language generation and text summarisation R&D has historically been conducted at several universities, including Lund and Gothenburg. At present, the most active research in this area is being pursued at Chalmers University of Technology (Gothenburg),[53] which has also resulted in a spin-off company, Digital Grammars,[54] and at Linköping University.[55]

Information retrieval and information extraction for Swedish are at present primarily being developed by commercial companies, for instance as parts of proprietary business intelligence and intranet search applications.

Licensing for LT tools developed by academic institutions is generally open-source, frequently in maximally permissive form, which allows commercial use as components of otherwise proprietary applications (e.g. the MIT license).

4.2 Projects, Initiatives, Stakeholders

Wallenberg AI, Autonomous Systems and Software Program (WASP) is the largest individual research program in Sweden aiming to "advance Sweden into an internationally recognised and leading position in the areas of artificial intelligence, autonomous systems and software".[56] The program supports projects that directly benefit LT, such as the building of state of the art Swedish deep-learning language models and improved speech technology algorithms.

The national research infrastructure *Nationella språkbanken*[57] 'the Swedish Language Bank'

[47] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner
[48] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface
[49] https://wasp-sweden.org/research/research-arenas/wara-media/
[50] See, e.g. https://translate.ling.helsinki.fi, https://blogs.helsinki.fi/found-in-translation/
[51] https://blogs.helsinki.fi/fiskmo-project/
[52] For instance, during the height of the COVID-19 pandemic, Swedish media reported more than once about the lack of translations – or perhaps even worse: the occurrence of mistranslations – of vital healthcare information into e.g. Somali and Arabic.
[53] http://www.cse.chalmers.se/research/group/Language-technology/
[54] https://www.digitalgrammars.com
[55] https://liu.se/en/article/lattlast-pa-webben
[56] https://wasp-sweden.org/about-us/
[57] https://www.sprakbanken.se. Nationella språkbanken is funded jointly by the Swedish Research Council and the participating institutions, at present through 2024.

collects, develops, manages and distributes LT resources for research. Organisationally, it is made up of the three divisions *Språkbanken Text*[58] (the text division, which also is a certified CLARIN B centre), *Språkbanken Tal*[59] (the speech division), and *Språkbanken Sam*[60] (the language and society division). It also administers *Swe-Clarin*,[61] the Swedish membership in CLARIN ERIC, in which an additional eight groups are involved nation-wide, both academic partners and public memory institutions (the National Library and the National Archives).

There is no dedicated national LT research funding program, but several LT-related data labs and some other LT-related projects have recently received national funding from Sweden's innovation agency, Vinnova.

On the consumer side, there is currently great interest in LT (or rather: language-centric AI) from both commercial enterprises and public agencies, and on the producer side, Sweden has a modest but thriving and growing spectrum of companies offering various LT and AI solutions. Both commercial enterprises and public agencies also develop in-house solutions, typically by hiring data scientists rather than LT specialists (at least in part because the latter are in short supply, due to lack of dedicated LT study programmes).

LT initiatives for societally central research

While it is clear that the main and most direct societal impact of LT comes in the form of language-aware devices and services, including services offered by public authorities and institutions, there are also more subtle and indirect ways in which LT will be increasingly important to our societies, and where it is unlikely that commercial providers will be forthcoming, which underscores the need for maintaining academic LT research that is not directly aiming at developing commercial or public-service applications.

One such salient societal aspect that is unlikely to attract much commercial support is accessibility. In ST in particular, applications range from hands-free and eyes-free speech interfaces and automatically read aloud texts to automatic subtitling and advanced hearing aids. Here, there is a distinct lack of resources for research that ensures that technology actually helps the user groups.

Further, an important contributing factor to the self-perception and cohesion of any community is an awareness and sense of a common history and cultural heritage. History is by definition studied through language in the form of texts and – over the last century – audio and video recordings. Knowledge of our history is important for present-day policy-making, and since both research and public awareness of history and cultural heritage are increasingly based on digital source materials, LT is increasingly needed to support this fundamental aspect of our societies.

Many Swedish memory institutions are digitising older texts and audio recordings, and academic reserch centers (but not industry) are building language resources out of these, as well as developing language tools for processing text corpora representing 800 years of Swedish, spanning several quite different historical stages of the language. An important role is played here by Swe-Clarin,[62] the Swedish node of CLARIN ERIC, which for instance has established a CLARIN knowledge center for diachronic language resources,[63] which has coordinated among others the design and compilation of a Swedish diachronic corpus, a reference dataset covering all stages of written Swedish from the middle ages (Old Swedish)

[58] https://spraakbanken.gu.se
[59] https://www.sprakbanken.se/omoss/organisationochverksamhet/sprakbankental.4.b86c4c173e68e512a37e3.html
[60] https://www.isof.se/lar-dig-mer/forskning/sprakbanken-sam
[61] https://sweclarin.se
[62] https://sweclarin.se/eng/home
[63] https://sweclarin.se/eng/centers/diares

onwards (Pettersson and Borin, 2019a,b,c).[64]

5 Cross-Language Comparison

The LT field[65] 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 (ELG) 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[66] 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

[64] https://cl.lingfil.uu.se/svediakorp/index_en.html
[65] This section has been provided by the editors.
[66] 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.

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 ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[67]

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[68] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the –currently in progress– development of a Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[69]

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.

[67] 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.

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

[69] 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.

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,[70] 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.

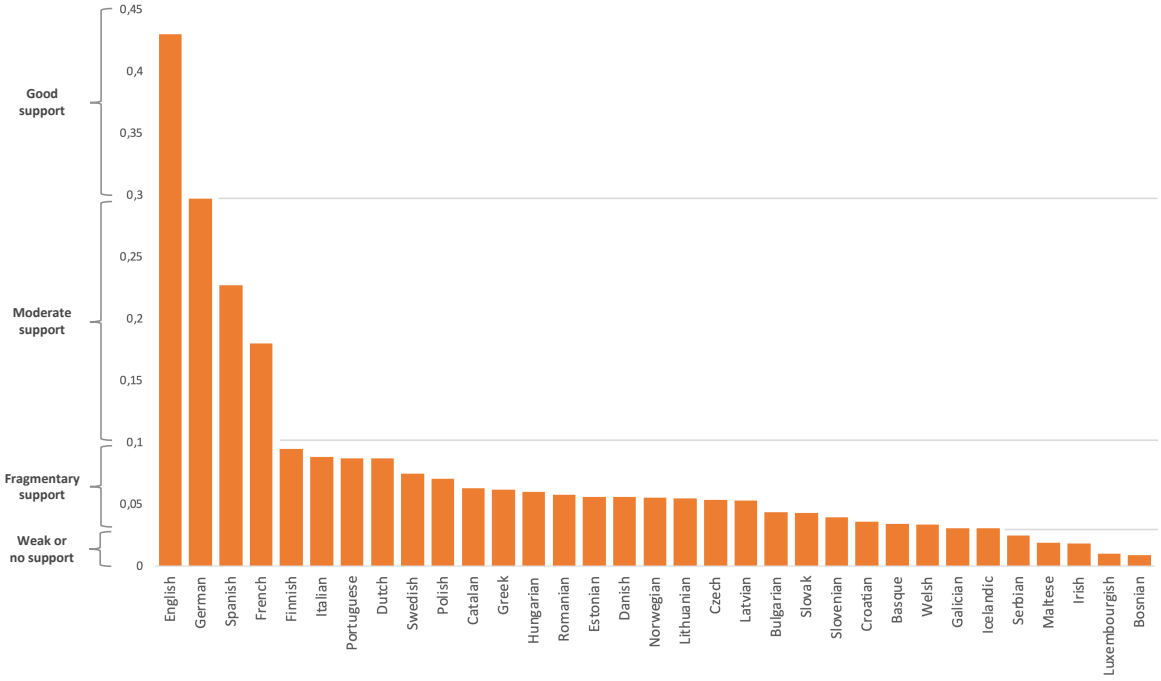

Figure 1: Overall state of technology support for selected European languages (2022)

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

[70] 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.

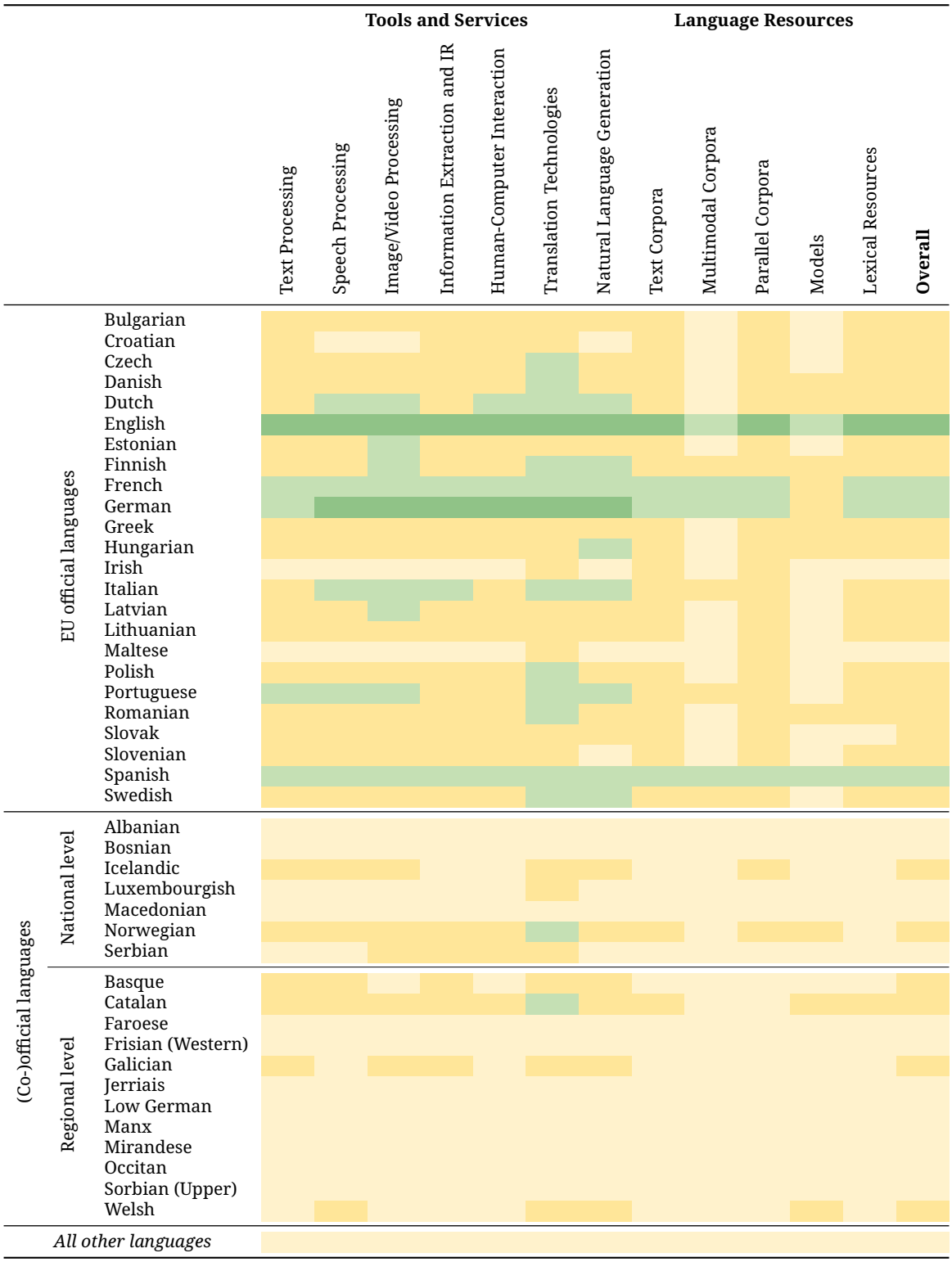

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

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

Sweden has a long history of academic LT R&D. The Speech Transmission Laboratory was established in 1951 and has – under various names – remained at the forefront of Swedish speech technology ever since, now hosting Språkbanken Tal. The computational linguistics research unit at the University of Gothenburg started its activities in the 1960s by collecting and processing the second modern large text corpus in the world (after the Brown corpus of American English),[71] and the precursor of Nationella språkbanken was established by this group in 1975, and it now forms Språkbanken Text.

A particularly important factor for the current state of Swedish LT was the *Swedish National Graduate School of Language Technology* (GSLT), which was run with government funding for roughly the first decade of the present millennium. This broad collaboration among all the academic institutions in Sweden where LT in some form could be pursued in PhD projects fostered a strong sense of community through joint national activities. During the decade of its existence, GSLT produced about 50 PhDs with an LT topic in academic disciplines ranging from library science to speech communication, but primarily from CL/NLP/LT.

For most of its existence, Swedish academic LT has been characterised by a well-balanced mix of researchers from computer science and linguistics (engineering and phonetics in the case of ST), and the undergraduate and masters programmes in LT have had the ambition to train their students in both areas. However, following the international trend (e.g. at ACL conferences), recent years have seen a clear shift towards LT researcher teams having a strong or pure computer science background, with an accompanying lack of awareness of

[71] This is *Press 65*, still available for searching and downloading through Språkbanken Text: https://spraakbanken.gu.se/korp/#?cqp=[]&corpus=press65, https://spraakbanken.gu.se/en/resources/press65.

many important linguistic aspects of LT research problems (e.g. Reiter, 2007; Wintner, 2009; Manning, 2015; Bender, 2016; Bender and Koller, 2020).

With the recent rise to prominence of deep learning LT systems (often under the guise of AI), some computer science centres have started showing an interest in LT as a central component of AI, more often than not without awareness of the long history and significant accomplishments of Swedish LT. In addition, both *commercial enterprises* and *public institutions*, other than universities, are showing an interest in developing language-aware applications for Swedish. In fact, as already mentioned, neural language model fine-tuning is becoming something of a cottage industry among Swedish public institutions (e.g., the Swedish Public Employment Service, the Swedish Tax Agency, and the Swedish National Financial Management Authority).

The Swedish academic LT expertise represents seventy years of accumulated knowledge, which should not be allowed to go to waste. Short-term, the best way of ensuring this is probably to focus on language resource development, where much work still remains to be done for Swedish. Well-designed gold-standard corpora for fine-tuning language models and evaluating LT systems will not emerge out of the blue, but require exactly this kind of expertise for their construction, not least in order to avoid pitfalls such as models making undesirable biased predictions that risk perpetuating gender roles or lead to unfair treatment of minority groups. Another major hurdle in this connection is presented by legal frameworks (e.g. IPR and GDPR), which in practice often effectively block research access to language data.

In the medium term, we should aspire to understand current language models – which typically come across as black boxes – in order to be able to exploit already existing linguistic knowledge (for instance, information about words collected in a lexical or conceptual resource) when training language models, which potentially will reduce their training data requirements, thus putting state of the art LT tools in reach of lower-resourced languages.

This calls for the establishment of closer collaborations and communication between the "traditional" LT research community and the new AI field, e.g., through dedicated LT training opportunities and earmarked funding for LT research.

References

Yvonne Adesam, Aleksandrs Berdicevskis, and Felix Morger. SwedishGLUE – Towards a Swedish test set for evaluating natural language understanding models. Technical report, University of Gothenburg, Gothenburg, 2020.

Lars Ahrenberg, Johan Frid, and Leif-Jöran Olsson. *A New Gold Standard for Swedish Named Entity Recognition: Annotation Guidelines*. Number SCRS-01-2020 in Swe-Clarin Report Series. Swe-Clarin, Online, 2020. URL https://sweclarin.se/sites/sweclarin.se/files/SCR-01-2020.pdf.

Gisle Andersen. Gjennomgang og evaluering av språkressurser fra NSTs konkursbo. Technical report, Språkrådet, Norway, Oslo, 2005. URL http://www.nb.no/sbfil/dok/dok.tar.gz.

Gisle Andersen. Akustiske databaser for svensk. Technical report, The Norwegian Language Bank, Oslo, 2011. URL https://www.nb.no/sbfil/dok/nst_taledat_se.pdf.

Emily M. Bender. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660, 2016.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL 2020*, pages 5185–5198, Online, 2020. ACL. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Aleksandrs Berdicevskis. The five lives of Talbanken. Online blog post, 2020. URL https://spraakbanken.gu.se/blogg/index.php/2020/06/09/the-five-lives-of-talbanken/.

Noam Chomsky. *Syntactic Structures*. Mouton, 1957. ISBN 978-90-279-3385-0.

Dana Dannélls, Lars Borin, and Karin Friberg Heppin, editors. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*. John Benjamins, Amsterdam, 2021. doi: 10.1075/nlp.14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL: HLT 2019, Volume 1*, pages 4171–4186, Minneapolis, 2019. ACL. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the World*. SIL International, Dallas, 24th edition, 2021. URL https://www.ethnologue.com.

Anna Hallberg. Kan man forska på riksdagsprotokollet. *Språkvård*, 1994(3):14–15, 1994. ISSN 0038-8440.

Einar Haugen and Lars Borin. Danish, Norwegian and Swedish. In Bernard Comrie, editor, *The World's Major Languages*, pages 127–150. Routledge, London, 3rd edition, 2018. ISBN 9781138184824.

Jens Ingman. Pts mobiltäcknings- och bredbandskartläggning 2020: en geografisk översikt av tillgången till bredband och mobiltelefoni i sverige. Technical Report PTS-ER-2021:16, Swedish Post and Telecom Authority (PTS), 2021.

Gunnar Ljusterdal. Språkvård i riksdagsprotokollet. *Språkvård*, 1973(2):5–9, 1973.

Christopher D. Manning. Last words: Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707, 2015.

Sjur Nørstebø Moshagen, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, and Per Langgård. European Language Equality. D1.42. Report on the Nordic minority languages. Technical report, European Commission, 2022.

Zimon Mossberg. Achieving automatic speech recognition for Swedish using the Kaldi toolkit. Ma thesis, KTH Royal Institute of Technology, Stockholm, 2016. URL http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-194178.

Norwegian Language Bank. Reorganized speech databases for Swedish ASR from Nordisk språkteknologi. Technical report, The Norwegian Language Bank, Oslo, 2020.

Robert Östling, Carl Börstell, and Lars Wallin. Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015)*, volume 23 of *NEALT Proceedings Series*, pages 263–268, Vilnius, 2015. NEALT. URL http://www.ep.liu.se/ecp/109/ecp15109.pdf.

Robert Östling, Carl Börstell, Moa Gärdenfors, and Mats Wirén. Universal dependencies for swedish sign language. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 303–308, Gothenburg, 2017. ACL. URL http://www.aclweb.org/anthology/W17-0243.

Mikael Parkvall. *Sveriges språk - vem talar vad och var?* Number 1 in Rappling. Institutionen för lingvistik, Stockholms universitet, 2009. URL http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-28743.

Mikael Parkvall. *Den nya mångfalden*. Number Det nya Sverige in Riksbankens Jubileumsfonds årsbok. Riksbankens jubileumsfond/Makadam Förlag, 2019. ISBN 978-91-7061-289-3.

Eva Pettersson and Lars Borin. *Characteristics of diachronic and historical corpora: Features to consider in a Swedish diachronic corpus*. Number SCRS-01-2019 in Swe-Clarin Report Series. Swe-Clarin, Online, 2019a. URL https://sweclarin.se/sites/sweclarin.se/files/SCR-01-2019_0.pdf.

Eva Pettersson and Lars Borin. *Swedish diachronic texts: Resources and user needs to consider in a Swedish diachronic corpus*. Number SCRS-02-2019 in Swe-Clarin Report Series. Swe-Clarin, Online, 2019b. URL https://sweclarin.se/sites/sweclarin.se/files/SCR-02-2019_0.pdf.

Eva Pettersson and Lars Borin. *Towards a Swedish diachronic corpus: Intended content, structure and format of version 1.0*. Number SCRS-03-2019 in Swe-Clarin Report Series. Swe-Clarin, Online, 2019c. URL https://sweclarin.se/sites/sweclarin.se/files/SCR-03-2019_0.pdf.

Christian Rauh and Jan Schwalbach. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. SocArxiv, 2020.

Manny Rayner, David Carter, Pierrette Bouillon, Vassilis Digalakis, and Mats Wirén. *The Spoken Language Translator*. Cambridge University Press, Cambridge, 2000. ISBN 978-0-521-77077-4.

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.

Ehud Reiter. Last words: The shrinking horizons of computational linguistics. *Computational Linguistics*, 33(2):283–287, 2007.

Maria Skeppstedt, Simon Dahlberg, Gunnar Eriksson, and Rickard Domeij. Texts and terms from Swedish public agencies in the SB Sam language bank. In *Proceedings of SLTC 2020*, Online, 2020. URL https://gubox.app.box.com/v/SLTC-2020-paper-5.

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

Niklas Vanhainen and Giampiero Salvi. Free acoustic and language models for large vocabulary continuous speech recognition in Swedish. In *Proceedings of LREC 2014*, pages 388–392, Reykjavik, 2014. ELRA. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/312_Paper.pdf.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, 2018. ACL. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Terje Wessel, Lena Magnusson Turner, and Viggo Nordvik. Population dynamics and ethnic geographies in Oslo: The impact of migration and natural demographic change on ethnic composition and segregation. *Journal of Housing and the Built Environment*, 33(4):789–805, 2018. ISSN 1566-4910. doi: 10.1007/s10901-017-9589-7. URL 10.1007/s10901-017-9589-7.

Shuly Wintner. Last words: What science underlies natural language engineering? *Computational Linguistics*, 35(4):641–644, 2009.