digitaliserings og ligestillingsministeriet

Folketingets Udvalg for Digitalisering og IT

Stormgade 2-6
1470 København K
Telefon 72 28 24 00
digmin@digmin.dk

Sagsnr.
2023-1967

Doknr.
20902

Dato
16-06-2023

Svar på spørgsmål fra Lisbeth Bech-Nielsen (SF) stillet den 22. maj 2023.

Spørgsmål nr. 77:

"Hvad er status på "Fælles Dansk Sprogressource", som beskrevet i VLAK-regeringens nationale strategi for kunstig intelligens (2019)? https://www.regeringen.dk/media/6537/ai-strategi_web.pdf".

Svar:

Der er indhentet svar fra Digitaliseringsstyrelsen, som oplyser følgende:

"I 2019 nedsatte Kulturministeriet et sprogteknologisk udvalg, som skulle undersøge status for dansk sprogteknologi, barrierer, der hæmmede udviklingen, samt give anbefalinger til indsatser, som kunne sætte gang i udviklingen af dansk sprogteknologi. Det sprogteknologiske udvalgs anbefalinger[1] blev skrevet ind i VLAK-regeringens "Nationale strategi for kunstig intelligens"[2] fra 2019, hvorved det offentlige forpligtede sig til at give dansk sprogteknologi et løft. Med baggrund i Digitaliseringspagten og Økonomiaftaler for 2020 med kommuner og regioner blev der igangsat et fællesoffentligt samarbejde for dansk sprogteknologi, under navnet sprogteknologi.dk, og som er finansieret til og med 2026.

Der har været en stor udvikling inden for kunstig intelligens og sprogteknologi den seneste tid. Et eksempel herpå er ChatGPT[3] og den underliggende sprogmodel, som har vakt stor opsigt siden offentliggørelsen i november 2022.

Der er dog en generel tendens til, at sprogteknologi for de mindre sprog som eksempelvis dansk udvikler sig væsentligt langsommere end for de større sprog og særligt engelsk.[4] Konsekvensen ved manglende dansk sprogteknologi er, at man ikke nyder samme anvendelsesmuligheder af ny teknologi på dansk som på de større sprog. I mange tilfælde betyder det, at man skal bruge sprogteknologi, som er trænet på engelsk til at løse opgaver på dansk. Den største udfordring for at udvikle sprogteknologi for de mindre sprog, herunder dansk, er, at der ikke eksisterer sprogressourcer af samme kvalitet og kvantitet, som fx på engelsk. Sprogressourcer er grundlæggende data, som repræsenterer sproget og kulturen, som algoritmer

[1] https://dsn.dk/wp-content/uploads/2021/01/sprogteknologi-i-verdensklasse.pdf
[2] https://digst.dk/media/19302/national_strategi_for_kunstig_intelligens_final.pdf
[3] Andre store sprogmodeller: Wu Dao 2.0., Megatron-Turing NLG, LEAM-1, OPT, Bloom, Luminous, Swtich Transformer, GlaM, PaLM (tre sidste er Googles)
[4] https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_9__Language_Report_Danish_.pdf

kan trænes på. Det kan være store mængder tekst, sammenhængende lyd og tekst eller sammenhængende billeder, lyd og tekst. Det danske sprogs særegne karakteristika, fonetisk reduktioner og ords flertydighed mv. gør det vanskeligt for maskiner at lære dansk.

I initiativet Fælles Dansk Sprogressource (i dag sprogteknologi.dk) arbejdes der med at styrke dansk sprogteknologi gennem blandt andet følgende centrale delprojekter:

1. Plaformen sprogteknologi.dk
2. Standarder for sprogressourcer
3. Et Centralt Ordregister
4. Et open source dansk taledatasæt
5. Udvikling af sprogressourcer på baggrund af eksisterende sprogdata
6. Europæisk og nordisk samarbejde
7. Deltagelse i andre projekter
8. Kommunikation, netværk og videndeling

I det følgende fremgår status på disse delprojekter.

1. Plaformen sprogteknologi.dk
Første led i indsatsen fokuserede på etableringen af webportalen sprogteknologi.dk, der samler og udstiller danske sprogressourcer på ét sted. Portalen udstiller nu metadata om 160 danske sprogressourcer, som frit kan anvendes til udvikling af dansk sprogteknologi.

2. Standarder for sprogressourcer
For anvendere, der skal arbejde med sprogdata, er det vigtigt, at formaterne er åbne og ens, simple, sammenlignelige og veldokumenterede. I en analyse identificerede Digitaliseringsstyrelsen brugernes behov, kortlagde eksisterende standarder på området og vurderede muligheden for genbrug, helt eller delvist, af eksisterende standarder. Med bidrag fra eksterne relevante parter er der udvalgt en række anbefalede standarder for sprogressourcer, som er publiceret på sprogteknologi.dk. Anbefalingerne blev publiceret på sprogteknologi.dk i marts 2022.

3. Et Centralt Ordregister
I marts 2021 blev udviklingen af et centralt ordregister (COR) igangsat, hvis formål er at udvikle og distribuere et register, der forsyner alle danske ord med et stabilt id-nummer og dermed gør det muligt for sprogmodeller at genkende danske ord fra hinanden, hvis de fx er enslydende eller staves ens. Udviklingsprojektet er et samarbejde mellem Digitaliseringsstyrelsen, Dansk Sprognævn, Det Danske Sprog- og Litteraturselskab og Center for Sprogteknologi ved Københavns Universitet. Registret vil blive frit tilgængeligt forventeligt ved udgangen af 2023, og der vil være adgang til løsningen via sprogteknologi.dk.

4. Et open source dansk taledatasæt
Digitaliseringsstyrelsen har indgået et bredt samarbejde mellem Alexandra Instituttet, Datalogisk Institut ved KU, Alvenir og Corti med henblik på at tilvejebringe et taledatasæt på dansk. Desuden har en række aktører, herunder ATP, givet tilsagn til projektet og skal deltage i projektets advisory board. Målet med et dansksproget talekorpus er, at fremtidige taleteknologier skal kunne genkende alle variationer af dansk, hvilket dermed skal sikre optimal brug og implementering af teknologien i samfundet. Endvidere kan taleteknologi bidrage til inklusion og ligestilling. Det færdigudviklede talekorpus blive stillet frit tilgængeligt for alle forventeligt ved udgangen af 2025.

5. Udvikling af sprogressourcer på baggrund af eksisterende sprogdata
Der eksisterer store mængder værdifulde sprogdata i mange offentlige institutioner. Eksempelvis findes der en række tekst- og lyddata hos eksempelvis DR, Det Kongelige Bibliotek, Rigsarkivet, Folketinget og Lex.dk. Disse sprogdata er af høj kvalitet, indeholder velforankret og nuanceret viden om dansk kultur og samfund og har derfor potentiale til at udgøre værdifulde sprogressourcer til udvikling af dansk sprogteknologi.

I dag er disse sprogdata enten ikke tilgængelige for eksterne anvendere, eller kun i et begrænset omfang, fx til forskning, og ellers er de ikke lagret i maskinlæsbart format, da de oprindeligt ikke er produceret med formål om at udgøre en sprogressource. Derudover er der en række udfordringer med at tilgængeliggøre sprogdataene pga. ophavsrettigheder og GDPR.

Digitaliseringsstyrelsen arbejder på at få tilgængeliggjort flere danske sprogdata. Pt. består arbejdet i at nedsætte en arbejdsgruppe, som skal undersøge de specifikke behov for sprogdata, herunder hvilke danske sprogdata der rådes over i arkiverne, juridiske og ressourcemæssige forhold, prioritering af vigtighed, behovet for efterbehandling mv.

Digitaliseringsstyrelsen arbejder også på at skabe domænespecifikke sprogressourcer, fx fra det finansielle-, bygge- eller sundhedsområdet, som er vigtige for, at sprogteknologiske løsninger kan tilpasses specifikke brugerbehov og erhverv, hvor fagtekniske begreber hyppigt anvendes. Digitaliseringsstyrelsen arbejder lige nu på et projekt, hvor der skal udarbejdes en sprogressource på baggrund af Region Hovedstadens offentlige dokumentsamling, som indeholder vejledninger til sundhedspersonale og borgere og indeholder derfor en række sundhedsfaglige begreber. Digitaliseringsstyrelsen vil sammen med bistand fra en forskningsinstitution samle dokumentsamlingen i et maskinlæsbart format og udgive den under en åben og fri licens. Det forventes, at der vil blive tilgængeliggjort flere sprogressourcer på baggrund af eksisterende sprogdata over de næste 1-3 år.

6. Europæisk og nordisk samarbejde

I EU arbejdes der i regi af den europæiske datastrategi henimod at etablere et såkaldt *language data space (LDS),* som skal etablere retningslinjer, governance og infrastruktur til at håndtere og dele sprogressourcer på tværs af EU. Etableringen af et europæisk *language data space* kan sikre, at private og offentlige aktører rentabelt kan udvikle og implementere sprogmodeller. LDS etableres forventeligt i 4. kvartal 2023.

Desuden er der på tværs af flere medlemsstater et igangværende arbejde med at rammesætte og etablere et såkaldt Europen Digital Infrastructure Consortium (EDIC) for sprogteknologi. EDICs er en helt ny mekanisme, der skal medvirke til at organisere flerlandeprojekter, og tanken bag den sprogteknologiske EDIC er bl.a. at bidrage til europæisk samarbejde om 1) tilvejebringelse af sprogdata gennem language data space og 2) udvikling af sprogteknologiske ressourcer og løsninger. Arbejdet er stadig i et tidligt stadie, og udkast til formelle beskrivelser af formål, finansiering og governance diskuteres på nuværende tidspunkt i en arbejdsgruppe bestående af repræsentanter fra medlemsstaterne. Arbejdet er igangsat i januar 2023, og det forventes, at ansøgningen om en permanent sprogteknologisk EDIC kan være klar i september 2023. Danmark deltager i dette arbejde.

Ligeså er en af de centrale ambitioner i Nordisk Ministerråd at fremme nordisk sprogteknologi, og forskning peger på, at der kan være synergier mellem sprogene ved at træne en stor fællesnordisk sprogmodel, da dette giver en bedre grundmodel, end hvis der trænes på hvert af sprogene for sig. Til et sådant nordisk samarbejde vil tilvejebringelsen af markant større danske sprogdatasæt, end der i dag er tilgængeligt, være nødvendigt. Endvidere råder Danmark ikke over tilstrækkeligt med computerkraft til at træne en stor dansk eller nordisk sprogmodel i stil med GPT-4. Der er derfor startet dialog med en række nordiske aktører for at undersøge mulighederne for et samarbejde om en fælles nordisk sprogmodel, da man eksempelvis i Sverige råder over tilstrækkeligt med computerkraft til at træne en fælles nordisk sprogmodel, og de har erfaringer med at udvikle store sprogmodeller.

7. Deltagelse i andre projekter

Digitaliseringsstyrelsen har i kraft af sprogteknologi.dk fulgt andre relevante projekter. Fx ved deltagelse i følgegruppen for signaturprojektet *"Digital inklusion og support ved talegenkendelse",* som arbejder med taleteknologi i forskellige services med fokus på digital inklusion i

Roskilde Kommune og Aarhus Kommune samt deltagelse i styregruppe for udrulning af digitale assistenter, som er et projekt drevet af Nota[5] under Kulturministeriet. Projektet går ud på at øge den digitale inklusion af ældre svagsynede og blinde ved brug af digitale assistenter.

8. Kommunikation, netværk og videndeling

Sprogteknologi.dk har en LinkedIn side af samme navn, som har eksisteret siden juni 2021. Det centrale omdrejningspunkt for indholdet på profilen er videndeling, troværdighed og genbrug af data og inspiration fra andres løsninger. Derudover afholder Digitaliseringsstyrelsen årligt Sprogteknologisk konference samt en række gå-hjem-møder, som har til formål at samle en bred skare af aktører fra de sprogteknologiske miljøer i Danmark for at skabe videndeling, inspiration og sparring om dansk sprogteknologi mellem det private, offentlige, forskningsinstitutioner og privatpersoner. Endeligt deltager Digitaliseringsstyrelsen i flere relevante eksterne konferencer for at indsamle nyeste viden om udviklingen inden for sprogdata- og løsninger."

Med venlig hilsen

Marie Bjerre

[5] Nota er et statsligt nationalbibliotek for mennesker med læsevanskeligheder, som er under Kulturministeriet. Nota producerer lydbøger, e-bøger og punktbøger til mennesker, der ikke kan læse almindelig trykt tekst. https://nota.dk/