DIGITALISERINGSSTYRELSEN

Evaluering af initiativet sprogteknologi.dk

Juni 2023

2023

Indhold

1. Introduktion

Sprogteknologi er et område, som i de seneste år har været i rivende udvikling, både på verdensplan og i Danmark. Det er én af de teknologier, der har størst potentiale for at gøre mange processer lettere og mere smidige – både for borgere og myndigheder. Det fællesoffentlige initiativ omhandlende sprogteknologi har til sigte at styrke dansk sprogteknologi og skal blandt andet understøtte danske virksomheder i udvikling af dansksproget kunstig intelligens (AI) og dermed bidrage til innovation og vækst målrettet det danske marked.

Digitaliseringsstyrelsen har i de seneste fire år været ansvarlige for initiativet for dansk sprogteknologi under titlen "sprogteknologi.dk". Initiativets indsats har i de første år taget udgangspunkt i det sprogteknologiske udvalgs anbefalinger anbefalinger (se bilag 1) Det sprogteknologiske udvalg blev nedsat af Kulturministeriet i 2019.

Imens har sprogteknologi generelt været i rivende global udvikling, hvilket også er kommet dansk sprogteknologi til gode. For eksempel er der internationalt blevet udviklet flere store sprogmodeller med milliarder af parametre, og der er kommet flere open source projekter, der udvikler værktøjer, bygger modeller, udstiller data og slutbrugerløsninger. Samlet giver det bedre udviklingsmuligheder for dansk sprogteknologi, og det bliver bredt diskuteret, hvorvidt Danmark skal have sin egen store sprogmodel. Dertil bliver løsninger, som involverer sprogteknologi, også i stigende grad anvendt i både den private og offentlige sektor i Danmark. Eksempler såsom tekstgenerering, chatbots og AI til hjælp i sagsbehandling har efterhånden bevist sit værd i det danske samfund. Situationen for dansk sprogteknologi er således anderledes end i 2019.

På den baggrund har Digitaliseringsstyrelsen afholdt to workshops, hvor interesserede i dansk sprogteknologi fik mulighed for at deltage. Det overordnede formål med workshoppen var at afdække relevansen af det sprogteknologiske udvalgs anbefalinger, samt diskutere initiativet sprogteknologi.dks hidtidige og fremtidige indsats. Indeværende rapport er en opsamling på de to workshops.

Rapporten er struktureret i syv afsnit. I det næste afsnit, Afsnit 2, præsenteres initiativet sprogteknologi.dk kort. Afsnit 3 gør rede for afholdelsen af workshoppen. Afsnit 4, 5 og 6 samler op på de drøftelser, der var på workshoppen. Til slut gør Afsnit 7 status på Digitaliseringsstyrelsens overvejelser for det videre arbejde med initiativet sprogteknologi.dk.

2. Initiativet sprogteknologi.dk

Dette afsnit indeholder en opsummering af initiativet sprogteknologi.dk, som blev vedtaget i 2019. Samarbejdet om et fællesoffentligt initiativ målrettet dansk sprogteknologi skal understøtte forskning inden for dansk sprogteknologi og støtte sprogteknologiske virksomheder i at udvikle dansksprogede løsninger inden for kunstig intelligens.

Danmark er et lille sprogområde med et komplekst sprog. Dansk har enslydende ord, hvor kun stødet adskiller dem fx 'hun' og 'hund', og mange ord har forskellig betydning og udtale, men staves på samme måde. Det gælder eksempelvis 'kost', der både er noget, man spiser, og noget, man fejer med. Samtidig findes der også ord der udtales ens, men staves forskelligt såsom 'vejr', 'vær', 'værd' og 'hver'. Og der findes også ord som 'så' der for eksempel både kan være datid af verbet 'se', det kan også betyde 'at så noget i jorden, og det kan bruges som adverbium – 'hun løb så hurtigt'.

Det gør det svært og dyrt at udvikle sprogteknologiske løsninger i forhold til markedspotentialet for den enkelte virksomhed. Hovedsageligt har det været de store internationale tech-virksomheder, som har drevet udviklingen inden for sprogteknologi generelt, men disse aktører har ikke dansk som en førsteprioritet, hvilket gør, at den danske brugerflade udvikles på baggrund af vilkårligt indsamlet sprogressourcer, da der oftest ikke afsættes de fornødne ressourcer til indsamling og udvikling af danske sprogressourcer. Derfor arbejder regeringen, KL og Danske Regioner med at understøtte udviklingen og samarbejdet med private leverandører, udviklere og forskere, så vi i fællesskab kan sikre, at teknologien tilpasses det danske sprog, således at teknologierne også fungerer på dansk. Initiativet har sigtet mod at samle relevante, eksisterende sprogressourcer og gøre disse frit tilgængelige digitalt, samt udvikle og udstille nye sprogressourcer, som kan mindske barrierene for mindre danske selskabers arbejde med udviklingen af dansk sprogteknologi.

Sprogteknologi.dk er netop udviklet som en central platform til samling og udvikling af sprogressourcer på dansk med henblik på at udstille disse til fri afbenyttelse for virksomheder og myndigheder, som arbejder med dansk sprogteknologi.

Initiativet har været tilrettelagt med et bredt og løbende samarbejde mellem de offentlige myndigheder, der anvender sprogteknologiske løsninger og det private markedet for udvikling af løsninger til formålet.

3. Workshops

I dette afsnit præsenteres workshoppernes rammer og hvordan de blev afholdt.

Digitaliseringsstyrelsen har afholdt to workshops af seks timers varighed hver. Workshopperne havde til formål at samle sprogteknologi.dks interessenter og give dem mulighed for at komme med feedback til initiativet, kommentere på udviklingen inden for dansk sprogteknologi, samt komme med forslag til fremtidige arbejdspunkter. Deltagerne blev inviteret af Digitaliseringsstyrelsen med formål om at samle forskellige arbejdsmæssige baggrunde for at få inputs fra forskellige perspektiver. Foruden dette blev der sendt en åben invitation ud på LinkedIn, så alle med interesse fik mulighed for at deltage.

De to workshops blev afholdt i februar og marts 2023, og fordelt på de to workshops deltog 35 personer. Heraf kom 11 deltagere fra offentlige organisationer, 20 deltagere kom fra den private sektor og 4 deltagere fra forskningsinstitutioner. Dagsordenen for workshoppen fremgår af bilag 2.

Hver workshop bestod af tre blokke med hvert sit fokuspunkt:

1) Evaluering af det forgangene arbejde i initiativet sprogteknologi.dk.
2) Kortlægning af eksisterende udfordringer for dansk sprogteknologi.
3) Fremadrettet arbejde.

Digitaliseringsstyrelsen faciliterede samtalerne og til hver af de tre fokuspunkter var et diskussionsoplæg. Som anslag til diskussionen fik deltagerne blandt andet udleveret et spørgeskema og senere blev de præsenteret for forskellige hypoteser samt mulighed for at kunne afgive deres anbefalinger til det fremadrettede arbejde.

4. Evaluering og status på initiativet sprogteknologi.dk

I dette afsnit præsenteres svarene på spørgeskemaerne og deltagernes evaluering af initiativet sprogteknologi.dk.

Evalueringen af initiativet har primært fokuseret på sprogteknologi.dks arbejdsmetoder, initiativets delprojekter og en diskussion af, hvorvidt initiativet har haft det rette fokus for at understøtte det danske sprogteknologiske miljø. Formålet med evalueringen har været at give Digitaliseringsstyrelsen indsigt i, hvilken værdi arbejdet i regi af initiativet har haft for interessenterne.
Hovedtemaerne for evalueringen var tre udvalgte temaer:

1) Platformen – sprogteknologi.dk
2) Nye danske sprogressourcer
3) Videndeling og samling af det sprogteknologiske miljø

Det er Digitaliseringsstyrelsens opfattelse, at interessenterne værdsætter det fællesoffentlige initiativ for dansk sprogteknologi. Dette ses både i det store fremmøde og interesse for workshoppen, de positive nedslagspunkter, der blev fremhævet til workshoppen, samt gennem interessenternes række af gode forslag til, hvordan det eksisterende initiativ kan forbedres og give yderligere værdi.

Platformen – sprogteknologi.dk

Platformen – sprogteknologi.dk – bliver generelt beskrevet som et værdifuldt værktøj for interessenterne, som primært bruger den til at skabe overblik over udviklingen af danske sprogressourcer. De ressourcer og nyheder, der bliver lagt op på hjemmesiden, giver god indsigt i udviklingen i dansk sprogteknologi. Ressourcerne på portalen bliver anvendt i nogen grad bl.a. til at træning af sprogmodeller – andre deltagere har tilkendegivet interesse i at benytte ressourcer til fremtidige projekter. Det fremkom dog også, at de internationale platforme, for eksempel Hugging Face, i højere grad bliver anvendt til at finde data, da sådanne sites bliver anset som industry standard.

Deltagerne gav en række forbedringsforslag til, hvordan portalen kan fungere som et offentligt vidensdelingssite med et par tilføjelser, for eksempel et showcase-katalog. Hertil blev der efterspurgt en række nye funktioner på portalen, for eksempel at ressourcer kan filtreres på baggrund af relevant domæne.

Nye danske sprogressourcer

I arbejdet med tilvejebringelse af nye danske sprogressourcer af høj kvaltiet, oplever workshopdeltagerne, at der er visse forbedringspunkter. Særligt blev det bemærket, at time-to-market er for lang, og kvaliteten af ressourcerne ikke stemmer overens med anvendernes behov. For eksempel nævnes det for lyd- og tekstdatasættet, som er udarbejdet med Nota[1], at lyden er for god, ensidig og ikke indeholder nok spontantale, samt at der er for lidt variation i stemmerne. Ovenstående indikerer, at lydbogsdata muligvis er svært at applicere i udviklingen af løsninger, der skal bruges i forskellige kontekster. Det skal dertil bemærkes, at data i Nota-datasættet er et frikøb af eksisterende lyddata for at teste nye måder at fremskaffe taledata på.

Som led i, at initiativet skal understøtte udviklingen af nye danske sprogressourcer, blev udviklingen af Det Centrale Ordregister (COR) igangsat i Q1 2021. Workshopdeltagerne har en række positive forventninger til dette delprojekt som forventes færdigt i december 2023. Allerede nu er indekset til COR samt manual og teknisk specifikation dog en realitet.[2] Det var dog kun et begrænset antal af deltagerne, som havde haft muligheden for at anvende COR inden workshoppen.

Videndeling og samling af det sprogteknologiske miljø

På workshopperne blev det tydeliggjort, at Digitaliseringsstyrelsens initiativ for at samle det sprogteknologiske miljø i Danmark har stor effekt. Samtlige deltagere hæftede sig særligt ved Sprogteknologisk konference, der afholdes sammen med Center for Sprogteknologi ved KU, som giver interessenterne mulighed for at koordinere, videndele og samle inspiration med hinanden. Der blev dertil efterspurgt flere lignende arrangementer, også gerne i mindre skala, som for eksempel gå-hjem-møder eller fokuserede netværksmøder for eksempel for én branche eller én type sprogteknologi.

Det blev nævnt, at initiativet sprogteknologi.dks formidlingsindsats er værdifuld, men oftest er målrettet et teknisk publikum. Sprogteknologi.dks LinkedIn showcaseside og nyheder på sprogteknologi.dk giver deltagerne mulighed for at følge udviklingen inden for dansk sprogteknologi. Der mangler dog nyhedsformidling, som henvender sig til et administrativt publikum, der ikke har indsigt i de tekniske detaljer omkring sprogteknologi.

Det vurderes således, at initiativet sprogteknologi.dk giver værdi, og at det bliver anvendt af forskere, kommunerne og private leverandører og udviklere til at orientere sig og koordinere på tværs af aktører i det sprogteknologiske miljø, men at det vil give værdi at udvide fokus ud over den sprogteknologiske kreds. Der er derfor plads til forbedringer, som kan løfte værdiskabelsen af initiativet yderligere.

---

[1] https://sprogteknologi.dk/dataset/notalyd-ogtekstdata
[2] https://ordregister.dk./

Dette vil Digitaliseringsstyrelsen tage til efterretning og arbejde videre med i det fremadrettede arbejde.

5. Barrierer for dansk sprogteknologi

I dette afsnit præsenteres barrierer for dansk sprogteknologi, der er identificeret på baggrund af inputs og observationer fra workshopdeltagerne. Inputtene er baseret på de konkrete erfaringer deltagerne har gjort i deres arbejde med sprogteknologiske projekter.

Digitaliseringsstyrelsen har i forbindelse med initiativet sprogteknologi.dk en række antagelser om, hvilke barrierer det sprogteknologiske miljø i Danmark står overfor. Derfor blev de to workshops også brugt til få afdækket, hvilke barrierer de deltagende interessenter møder i forbindelse med at udvikle danske sprogteknologiske løsninger. Som en del af afdækningen blev deltagerne spurgt ind til, hvem de forventer, der skal løse udfordringerne i fremtiden.

Juridiske barrierer

Den mest hyppigt nævnte barriere blandt workshopdeltagerne er de juridiske barrierer. Enten fordi juridiske restriktioner er en hæmsko, eller fordi relevante juridiske regler er uprøvet på området, og derfor er det uklart, hvordan juraen skal tolkes. Særligt GDPR fremhæves som værende et aspekt, der tærer på de økonomiske og tidsmæssige ressourcer i et udviklingsprojekt, der involverer dansk sprogteknologi. Denne problemstilling ses også i den generelle opsamling fra "Signaturprojekterne".[3] Til workshoppen blev den juridiske barriere yderligere uddybet med, at ikke to udviklingsprojekter er ens, hvorfor det juridiske arbejde i mange tilfælde skal starte fra begyndelsen. Det vanskeliggør at lave en altomfattende indsats, der lemper barrieren i bred forstand.

Der var dog enighed blandt deltagerne om, at initiativet sprogteknologi.dk kan gøre en forskel for det sprogteknologiske miljø i Danmark. Herunder for eksempel ved at skabe klarhed omkring de juridiske rammer, hjælpe interessenterne med at besvare juridiske spørgsmål eller komme med generelle anbefalinger/retningslinjer, som interessenterne kan forholde sig til, eksempelvis ved at fremhæve gode juridiske show-cases.

Manglende teknologiforståelse blandt beslutningstagere

Deltagerne fremhævede manglende forståelse for sprogteknologi, data, dataadgang og tekniske løsninger blandt ledere og beslutningstagere som en barriere. Det

[3] For år 2021 https://digst.dk/nyheder/nyhedsarkiv/2021/april/digitaliseringsstyrelsen-kl-og-danske-regioner-opsamler-erfaringer-og-udfordringer-med-kunstig-intelligens-i-det-offentlige/ og for år 2022 https://digst.dk/nyheder/nyhedsarkiv/2023/februar/statusrapport-tager-mellemtid-paa-arbejdet-med-kunstig-intelligens-i-signaturprojekterne-i-kommuner-og-regioner/

skyldes særligt, at teknologiens anvendelsesmuligheder og potentialer fortsat udforskes, og der mangler konkrete eksempler på fordelene ved teknologien, herunder særligt eksempler der belyser konkrete anvendelsesscenarier og gevinsterne herved.

Der var overordnet enighed om, at det er en fælles opgave at skabe forståelse for dansk sprogteknologi blandt et større publikum. Det er deltagernes opfattelse, at Digitaliseringsstyrelsen og offentlige myndigheder har et ansvar for at få skruet op for de politiske ambitioner på de højere politiske niveauer, udbrede kendskabet til effektive løsninger med dansk sprogteknologi og på den måde få potentialet og udfordringerne frem i lyset. Udfordringen for private udbydere, leverandører m.m. ligger i manglende indsigt i kundernes forretning og praksis. Mange af de private leverandører har svært ved at identificerer kundernes behov, primært for dem der ikke normalt beskæftiger sig med sprogteknologi. Særligt blev der udtalt et behov for et showcase-katalog med inspiration til udviklingsprojekter, som kan give det sprogteknologiske miljø et arsenal af eksempler, som kan benyttes til at eksemplificere teknologiens kunnen over for ledere og beslutningstagere. Et sådant katalog kan 'bo' på portalen sprogteknologi.dk.

Datamangel

Flere af deltagerne gav udtryk for, at der er kommet flere danske tekstdata på forskellige domæner. Dog efterlyses der stadig mere data, både generelt, men også inden for specifikke domæner såsom medicinsk tekst. Det blev også foreslået, at man kunne kigge ind i juridiske løsninger for at tilgængeliggøre data, som ikke er offentligt tilgængelige af forskellige årsager. For eksempel ved at træne på data i lukkede miljøer, anonymiserede data eller syntetiske datasæt. Desuden søgte deltagerne flere annoterede (trænings)datasæt generelt af både høj kvalitet, men også større mængder.
Deltagerne meddelte endvidere, at mange af de efterspurgte data eksisterer, men at adgangen til disse er vanskelig, samt mulighederne for at benytte f.eks. journaldata indebærer et ressourcekrævende stykke juridisk arbejde. Ift. til lyddata er der fortsat mangel på tilgængelige ressourcer af rette kvalitet. Som oftest har virksomhederne kun adgang til lyddata via de organisationer, de udvikler en løsning til, og det er derfor ikke noget, alle aktører kan få adgang til.

For at etablere en form for målbarhed i kvalitet af datasæt og sprogmodellers performance på dansk er der ligeledes behov for evalueringsdatasæt eller benchmarks til test af kvalitet af datasæt, modellers forståelse for dansk kultur osv. Evalueringsdatasæt anvendes til løbende at kvalitetssikre og dermed forbedre de sprogmodeller, der anvendes i produkter og services.

Mangel på flere specialister

En fjerde barriere, der blev pointeret, er manglende specialistviden inden for dansk sprogteknologi. For at udvikle dansk sprogteknologi af høj kvalitet kræver det specialistviden omkring det danske sprog, samt teknisk forståelse for kunstig intelligens, herunder særligt sprogteknologi. Fx vil en professionel, der ikke taler

dansk, have sværere ved at finde sproglige fejl eller manglende kulturelle nuancer i de tekniske løsninger. Med til dette punkt, blev det fremhævet, at det kræver uddannelse af flere studerende inden for feltet, om end der ikke eksisterer mange eller nok uddannelser på danske universiteter, der specifikt beskæftiger sig med dansk sprogteknologi, både hvad angår udvikling, implementering og praksis. De uddannelser, der har et sådan fokus, har kun få pladser til et stigende antal ansøgere og beskæftiger sig kun med udviklingen af sprogteknologi. Specialister har i højere grad mulighed for at udfordre udviklingen og sætte spørgsmålstegn ved nye teknologiske trends såsom Chat-GPT og rådgive omkring, hvorvidt det er nødvendigt med en Chat-GPT-DK samt implikationerne af sådan en løsning. Der er lige nu kun få eksperter inden for disse områder, som dog alle yder en stor indsats for at fremme dansk sprogteknologi.

Mangel på kvalificeret arbejdskraft er også en udfordring for arbejdsgivere, som kun modtager få kvalificerede ansøgere til stillinger inden for natural language processing (NLP).

Manglende infrastruktur

Mangel på hardware og computerkraft i Danmark var en af de gennemgående temaer på de to workshops. Vi besidder ikke tilstrækkelig stor computerkraft i Danmark til at træne en stor, åben dansk grundmodel, og virksomheder, myndigheder og forskere baserer derfor ofte deres finetunede sprogmodeller på en udenlandsk grundmodel. Træning af en stor, åben dansk sprogmodel vil gavne udviklingen af dansk sprogteknologi, da den vil give øget transparens i, hvilke data der trænes på samt sikre, at disse data overholder juridiske regler som GDPR og ophavsret. Ved udvikling af en stor dansk sprogmodel vil der være fokus på, at modellen er trænet på data, der afspejler de danske værdier, og at det danske sprog og vendinger afspejles mere nuanceret end i de udenlandske sprogmodeller. Endvidere vil det i højere grad være gennemsigtigt for anvendere, hvilke bias og svagheder modellen besidder i forhold til dansk kultur og ligestilling.

På workshoppen blev der afholdt et oplæg fra Nvidia om GPT-SW3 og om, hvordan man i Sverige har fået doneret en kraftfuld computer, som har gjort AISweden i stand til at træne GPT-SW3. AISweden har nævnt, at de gerne vil inkludere dansk træningsdata i deres model, så modellen bliver fællesnordisk og derved ligeledes kan anvendes open source som en stor dansk grundmodel. Dog er der juridiske barrierer, der står i vejen for, at dansk træningsdata kan overdages til AISweden.

6. Fremadrettede anbefalinger for dansk sprogteknologi

I dette afsnit præsenteres anbefalinger for ønsker til initiativets fremtidige arbejde.

Digitaliseringsstyrelsen søger at arbejde ud fra interessenternes behov. Derfor fik workshoppernes deltagere mulighed for at afgive forslag til, hvordan Digitaliseringsstyrelsen fremadrettet skal fokusere arbejdet med det fællesoffentlige initiativ for dansk sprogteknologi. De er listet herunder efter antal tilkendegivelser.

Nordisk samarbejde

Der er et udbredt ønske om et øget samarbejde på tværs af de nordiske lande. Dette kan være i form af datadeling, udvikling af sprogmodeller, samt større politisk fokus på sprogteknologi fra de skandinaviske politikere. Digitaliseringsstyrelsen kan have en faciliterende rolle i at tilvejebringe et sådant samarbejde angående nordisk sprogteknologi. Mulige resultater af et øget nordisk samarbejde kan være en fælles nordisk GPT-model eller bedre videns- og erfaringsudveksling landende imellem.

Udbredelsesstrategi generelt

Dette punkt dækker over flere ønsker. Det ene er, at kendskabet til potentialet af sprogteknologi hos eksperter, politikere og borgere skal øges. Der blev foreslået konkrete måder, dette kan gøres på, bl.a.: showcases, som kan være i form af domænespecifikke blogindlæg, som kan udstilles på sprogteknologi.dk-platformen, og som løbende opdateres. Disse cases kan også fremvises ved meet-ups og/eller netværk. Derudover kan der udarbejdes handlepunkter, som skal udføres af enten Digitaliseringsstyrelsen eller eksterne aktører, for at synliggøre processen og gøre det til et fælles initiativ. Til udbredelsesstrategien ligger også et ønske om bedre kommunikation fra styrelsens side, som kan ramme flere ikke-tekniske aktører.

Policy

Deltagerne peger på, at der ligger et arbejde med at få sprogteknologi på dagsordenen hos politikerne. Dette kunne f.eks. være at undersøge mulighederne for pligtaflevering af data til sprogteknologiske formål ved offentlige udbud eller "at tænke på sprogteknologi som infrastruktur" og at have "dansk machine learning på Finansloven". Et håndgribeligt finansieringsprojekt er f.eks. en stor super computer til at træne sprogmodeller på, som man ser det i Sverige og deres pre-release af en stor sprogmodel GPT-SW3.
En anden udfordring ligger i, at der er for lidt hjælp og rådgivning at hente i arbejdet med GDPR. Der ligger en stor arbejdsbyrde for mange i at sætte sig ind i

GDPR-regler og good-practices for, hvordan man overholder GDPR og andet gældende lovgivning, når man skal have adgang til data, behandle data med flere. Der vil blive arbejdet på at oprette løsninger til at lette juridiske udfordringer.

Dataadgang

Der er efterspørgsel efter at få gjort noget ved adgang til data. Som nævnt kan et muligt greb være at få indsat regler for pligtaflevering af sproglige data ved vundne offentlige udbud til enten Rigsarkivet eller det Det Kongelige Bibliotek, så man fremadrettet kan sikre tilførsel af relevante sprogdata. Generelt findes der meget data i Danmark, men adgang til det er ofte ikke tilgængeligt til at træne sprogteknologi pga. GDPR, ophavsrettigheder eller formatet. Steder at starte kunne være at undersøge mulighederne for at Det Kongelige Bibliotek, Rigsarkivet med flere kan åbne for mere fri sprogdata til træning af sprogmodeller.

Benchmark datasæt

For at kunne måle status quo og kvantificere fremgang ved arbejdet med dansk sprogteknologi og især store sprogmodellers performance på dansk, blev det foreslået at udvikle evaluerings- og benchmark datasæt og domænespecifikke evalueringsstandarder. Dette kan være datasæt, som kan være med til at vurdere en sprogmodels kendskab til forskellige danske kulturelle fænomener såsom Bamse og Kylling i stedet for at beskrive et typisk børneprogram som det amerikanske Sesame Street, eller til at vurdere, hvor god en sprogmodel er til at forstå danske ordsprog eller dialekter.
Sådanne datasæt vil skabe gennemsigtighed for mulige købere af sprogteknologi, da de lettere vil kunne se kvaliteten af den model, som bliver anvendt i et givent produkt. Det vil gøre det lettere for forskere og udviklere at undersøge, hvordan tilpasninger eller ændringer i træningsformen har konsekvenser/betydning for en sprogmodel. Slutteligt vil et benchmark datasæt skabe grobund for en konkurrence i det sprogteknologiske miljø om at skabe den bedste model på markedet.

Fokus på uddannelser

I Danmark mangler vi arbejdskraft inden for it-domænet, og dette gør sig også gældende inden for dansk sprogteknologi. I den nyligt udgivne Redegørelse om Danmarks digitale vækst[4] er en af hovedkonklusionerne, at over 60% af danske virksomheder har svært ved at få fat i it-specialister, og at tallet har været stødt stigende de seneste år.
Desuden er der i skrivende stund kun få uddannelser som kan opbygge relevante kompetencer inden for arbejdet med dansk sprogteknologi og disse har få pladser. Det sprogteknologiske udvalgs langsigtede anbefaling om en styrkelse af kompetencer og uddannelse inden for dansk sprogteknologi vurderes derfor stadig som et område, hvor der kan gøres en større indsats.

---

[4] DIU Almdel Bilag 75 Redegørelse om Danmarks Digitale Vækst 2023pdf (ft.dk)

7. Konklusion og fremadrettet arbejde

I dette afsnit sammenfattes konklusionerne for denne rapport og der bliver sat nogle hegnspæle for, hvilke områder Digitaliseringsstyrelsen kommer til at arbejde videre med i fremtiden.

Sprogteknologi har rykket sig utroligt meget de seneste fire år. Kendskabet til sprogteknologi, potentialer såvel som problemstillinger er steget hos den almene borger.

Digitaliseringsstyrelsen har blandt andet bidraget til udviklingen ved at samle og koordinere på tværs af lande og brancher med de sprogteknologiske konferencer i samarbejde med Center for Sprogteknologi ved KU. Sådanne tiltag er noget interessenterne gerne ser mere af i mange formelle og mindre formelle afskygninger. De to afholdte workshops har vist, at anbefalingerne, der er arbejdet ud fra, har brug for en opdatering, og at Digitaliseringsstyrelsen også fremadrettet arbejder agilt.
Det står klart, at det arbejde Digitaliseringsstyrelsen har lavet i forbindelse med initiativet sprogteknologi.dk har været overvejende positivt. Platformen bliver brugt til at dele viden og til at orientere sig om udviklingen af ny dansk sprogteknologi. Den kunne dog gentænkes til at indeholde nye funktioner for at afhjælpe nogle af de nævnte barrierer.

Digitaliseringsstyrelsen har løbende tilvejebragt nye danske sprogressourcer såsom frikøb af lyddata samt nedsat nogle anbefalede standarder for formatering og genanvendelighed af sproglige data. Her har konklusionen været at time-to-market har været lang på nogle af leverancerne, selvom produkterne fortsat er efterspurgt.

Der er fortsat behov for et fællesoffentligt initiativ for dansk sprogteknologi, da der fortsat er store udfordringer for de sprogteknologiske aktører. GDPR og andre juridiske barrierer er blandt de største udfordringer. Deltagerne påpegede, at det vil være særdeles værdiskabende, såfremt de kan få hjælp til håndtering af disse udfordringer, og at Digitaliseringsstyrelsen kan spille en vigtig rolle heri.

Det er svært at komme igennem med budskaber om potentialer og udfordringer til beslutningstagere og politikere, da der mangler de gode historier og business-cases på dansk, som kan formidles til både offentlige myndigheder og private virksomheder. Også her ses der en mulighed for, at Digitaliseringsstyrelsen kan bidrage ved at kommunikere flere show-cases, fx via sprogteknologi.dk, Sprogteknologisk konference, LinkedIn og gå-hjem-møder mv.

Denne udfordring har også tråde til den manglende uddannelse af folk med de rigtige kompetencer inden for sprogteknologi til at kunne bevæge sig i spændfeltet mellem teknologier og mennesker og netop formidle disse cases.

Mængden af tilgængelige danske sprogressourcer er fortsat begrænset, og der mangler derfor sprogressourcer til udvikling af state of the art dansk sprogteknologi. Utilgængelighed af data er stadig en udfordring, både i store mængder, men også inden for mere domænespecifikke områder, og det samme er annoterede datasæt til træning af sprogmodeller og muligheden for at evaluere disse. Disse annoterede datasæt er ikke billige at skabe, og det er en hæmsko for mange aktører og projekter. Dertil udgør manglen på computerkraft også en barriere, fx for udviklingen af en stor dansk sprogmodel.

Fremadrettet ser deltagerne gerne en øget eller fortsat indsats fra Digitaliseringsstyrelsen. Der er blevet kigget meget til de andre skandinaviske lande, deres tiltag og udviklingsprojekter. Det er efterhånden bredt diskuteret, hvorvidt man i Danmark skal starte med at udvikle en dansk GPT model, på samme måde som man for eksempel har gjort det i Sverige. Meget tyder derfor på, at den generelle stemning i Danmark peger i retningen af et øget nordisk samarbejde. Der ligger også en stor opgave i policyarbejdet angående afhjælpningen af juridiske barrierer, hvilket er noget Digitaliseringsstyrelsen kan give sig i kast med for at gøre hverdagen nemmere for mange af deltagerne.
Der er fortsat stor opbakning til, at Digitaliseringsstyrelsen afholder flere samlinger og events for at samle og udbrede viden.

8. Bilag og deltagerliste

Heri findes bilag, samt oversigt over de deltagende organisationer fra de to workshops

Bilag 1:

'Sprogteknologisk udvalg anbefaler:

1. Oprettelse af en organisation med ansvar for at etablere en dansk sprogbank og for at planlægge og igangsatte sprogteknologiske udviklingsprojekter
2. Sprogbanken skal tilvejebringe og vedligeholde danske sprogresurser som skal stilles til rådighed i høj lingvistisk kvalitet optimeret til sprogteknologiske formal.
   Sprogbanken skal som minimum indeholde:
   2.1. Et tidskodet dansk talesprogskorpus
   2.2. En sprogteknologisk værktøjskasse
   2.3. Danske tekstkorpusser og opmærkede guldstandarder
   2.4. En avanceret dansk orddatabase
   2.5. En dansk termbank
   2.6. En resurseportal til distribution og deling af sprogresurser i sprogbanken
3. Styrkelse af kompetenceudvikling og uddannelser inden for dansk sprogteknologi
4. Styrkelse af forskning i dansk sprogteknologi.'

Kilde: Sprogteknologi i verdensklasse

Bilag 2:

Tid	Programpunkt
09:45 – 11:20	Første emne: Evaluering af initiativet for dansk sprogteknologi
11:20 – 12:00	Frokost
12:00 – 12:50	Andet emne: Barrierer for dansk sprogteknologi
12:50 – 13:00	Pause
13:00 – 14:00	Præsentation GPT-SW3 af Nvidia
14:00 – 14:10	Pause
14:10 – 15:00	Tredje emne: Inspiration, potentialer og muligheder
15:00 – 16:00	Netværkskaffe

Deltagende organisationer:

Aarhus Universitet
Alexandra Instituttet
Alvenir
ATP
Capturi
Center for Sprogteknologi på KU
Copenhagen Institute of Future Studies
Dansk Sprognævn
Det Danske Sprog- og Litteraturselskab
DI Digital
Dictus
EFNIL
Ekstra Bladet
KL
Københavns Universitet – Institut for Engelsk, Germansk og Romansk
MediaCatch
Mirsk
Nightingale.io
Nvidia
PWC
Omilon
Roskilde Kommune

Sønderborg Kommune
Vitec