

Dansk AI

A. INDLEDENDE OPLYSNINGER

Aktivetsområde	Indsatsområdet Digital sikkerhed, tillid og dataetik
Institut	Alexandra Instituttet
Titel <i>Dækker indholdet af aktiviteterne</i>	Dansk AI
Nummerering <i>Af beskrivelsen</i>	4
Version	1
Periode <i>Forventet start og slut</i>	1/1 2024-31/12 2024
Kontaktperson	Kristian Krämer

B. ÆNDRINGER

Angiv her, hvis en planlagt aktivitet er ændret i forhold til den forudgående version af beskrivelsen.

C. BESKRIVELSE

1. Mål <i>Hvorfor? Hvad er målet for aktiviteterne? Hvordan bidrager de til det overordnede mål for indsatsområdet?</i>	Målet med aktiviteten er at styrke danske virksomheders muligheder for at udvikle og benytte software, der kan håndtere det danske sprog. Både for at styrke deres konkurrenceevne, men også for at danske forbrugere og borgere kan få gavn af software, der kan forstå det danske sprog. Det kan både være en chatbot der guider en igennem et køb i en webshop eller software der kan transskribere telefonsamtaler, så det er hurtigere at lave journalisering, notater, m.v. i borgerservice. Udviklingen indenfor sprogteknologi går lynhurtigt i disse år, men mindre sprog som dansk sakker forsat bagud. Det betyder at trods den tiltagende opmærksomhed på dansk sprogteknologi de seneste år, er dansk fortsat et begrænset sprog digitalt set. Konklusionen fra den seneste rapport om Europæiske sprogressourcer er meget klar. Udviklingen går i positiv retning, men dansk betragtes forsat som "fragmentarisk understøttet" og vi er i Danmark forsat et godt stykke efter de større europæiske sprog som engelsk, tysk og spansk indenfor alle sprogteknologiske områder.¹ Moderne sprogteknologiske løsninger baserer sig på maskinlæring og det kræver store mængder af data i høj kvalitet. Og de data mangler vi på dansk for at kunne følge med den udvikling, der sker på de større sprog. Set fra et virksomhedsperspektiv er mangel på data forsat en af de mest kritiske udfordringer. Da data er omkostningsfulde at producere, er det en så betydelig barriere, at mange virksomheder afventer at gå i gang. Der er med andre ord brug for at udvikle åbne danske datasæt, som kan bruges af alle. Åbne data vil desuden betyde at de store internationale og ofte flersprogede grundmodeller, der ofte udvikles af globale tech-virksomheder vil kunne inkludere disse data i deres træning af modeller, så de også bliver bedre til dansk (se fx Open AI's <u>Whisper</u>-model).
---	---

¹ https://ec.europa.eu/info/sites/default/files/about_the_european_commission/service_standards_and_principles/documents/elis2022-report.pdf

	Hvis ikke vi gør noget bliver det kun sværere og sværere at konkurrere med de større sprog, og dermed risikerer vi at dansk 1) bliver et irrelevant marked for sprogteknologiske løsninger og 2) at dansk bliver et irrelevant sprog digitalt. Konkret bidrager der i aktiviteten til de overordnede mål-indikatorer med et eller flere case forløb sammen med virksomheder, videreudvikling af teknologisk service indenfor AI, kompetenceopbygning, udbygning af samarbejde med videnspartnere, samt vidensspredning af resultater til dansk erhvervsliv og andre interesserede i form af f.eks. indlæg på konferencer, webinarer, formidlingsrapporter, artikler, blogindlæg, m.v.
2. Indhold Hvad skal der ske? Hvilke(n) konkret(e) aktiviteter udføres?	Med udgangspunkt i ovenstående målbeskrivelse gennemføres følgende aktiviteter i 2024: Udvikling af danske tekst- og tale datasæt og modeller: Denne delaktivitet fokuserer på tre primære områder, hvoraf to har særlig opmærksomhed på taledata og -teknologier. Dette skyldes, at arbejdet i både 2022 og 2023 har vist, at dette område er særligt udfordrende og komplekst for virksomheder at håndtere selv. <i>Produktion af danske taledata:</i> Der har længe været efterspørgsel fra danske virksomheder efter et omfattende, annoteret dansk taledatasæt. Dette skyldes både den begrænsede tilgængelighed af data i dag, udfordringer ved brug af egne data på grund af GDPR og de høje omkostninger ved produktion af taledata. I 2024 vil fokus være på at producere danske taledata med bred repræsentation af talt dansk, samtidig med at der minimeres uhensigtsmæssig bias mod dialekter og sprogstile. Denne delaktivitet overlapper med projektet CoRal (se nedenfor), og der vil desuden blive arbejdet på at etablere yderligere synergiprojekter for at øge den samlede mængde af tilgængelige taledata, da dette er en væsentlig mangelvare hos virksomheder. <i>Udvikling af danske talemødder:</i> I 2024 vil der blive udviklet forskellige talemødder, herunder tale til tekst, tekst til tale, akustiske mødder og domænespecifikke mødder. Disse mødder vil blive udviklet og afprøvet i samarbejde med virksomheder og/eller offentlige organisationer for at illustrere anvendelsen af taleteknologi på tværs af forskellige industrier. Der vil også eksperimenteres med at udvikle destillerede mødder - komprimerede versioner af de nævnte mødder, der kræver mindre computerkraft og energi at afvikle. Dette kan være en betydelig fordel for virksomheder, der ønsker at eksperimentere med forskellige taleteknologiske mødder i deres nuværende løsninger eller teste potentielle use cases. <i>Udvikling af danske LLMs (Large Language Models):</i> Store sprogmodeller, også kendt som Large Language Models (LLMs), udvikles ved at træne på omfattende mængder ikke-annoteret dansk tekst. Disse mødder danner grundlag for specialiserede sprogmodeller, der kan udføre forskellige opgaver, såsom følelsesanalyse eller opdagelse af upassende sprogbrug. For at tilpasse grundmodellen til en specifik opgave, finjusteres den ved at tilføje et mindre datasæt, der repræsenterer opgaven. De danske LLMs spiller en afgørende rolle for at sikre kvaliteten af de specialiserede mødder, både med hensyn til præcision og bekæmpelse af eventuelle uønskede bias. De mest anvendte LLM'er, som f.eks. GPT-3.5, fungerer godt på dansk, men på grund af GDPR-udfordringer er de ikke ideelle til brug i den offentlige sektor eller på personfølsomme data. Derfor er der behov for at udvikle nye danske mødder, der kan anvendes uden at overtræde GDPR. Det er vigtigt at bemærke, at træningen af danske LLMs er en ressourcekrævende proces, der kræver betydelig computerkraft og har en indvirkning på miljøet. Derfor er det en fornuftig beslutning at gøre disse mødder åbent tilgængelige for at sikre, at de kan gavne alle, der har behov for dem.

	Udvikling og lancering af AI TDU. I 2024 lancerer Alexandra Institutet TDU'en for AI, som vil omfatte en bred vifte af temaer og tilgange til AI. Denne aktivitet vil bidrage med rådgivning inden for dansk data science med fokus på tekst- og taledata. Formålet med denne delaktivitet er at løfte den nuværende TDU fra en prototype til en fuldt operationel version. TDU'en henvender sig specifikt til machine learning-udviklere i danske virksomheder, der allerede arbejder med dansk data science eller ønsker at starte. TDU'en tilbyder end-to-end rådgivning og udvikling af softwareløsninger baseret på tekst- og taledata inden for sprogteknologi med særlig fokus på dansk. Derudover indeholder TDU'en tekniske funktioner, der støtter danske virksomheder i at (1) få overblik over tilgængelige open source-muligheder og træffe det bedste tekniske valg samt (2) adressere og minimere udfordringer, som udviklere ofte står over for. Målet er at lette udviklingen, opdateringen og vedligeholdelsen af maskinlæringsbaserede systemer, hvilket i nogle tilfælde kan være afgørende for en positiv business case. Alle funktioner vurderes og justeres løbende i samarbejde med markedsaktører, følgegruppen og andre interessenter for at sikre, at der ikke udvikles funktioner, der skaber markedsforvridning. Udvikling af samarbejde med videnspartnere Samarbejdet med både DIKU og AU fra de foregående år fortsættes og udbygges i 2024, med særlig vægt på udvikling af taledata og talemødelere. En mindre del af budgettet for aktiviteten vil blive allokert til udvikling af synergiprojekter og ansøgningsskrivning. I forbindelse med TrustLLM-projektet udvides samarbejdet yderligere med internationale partnere, herunder Linköpings Universitet (LIU), Fraunhofer, Forschungszentrum Jülich, Lindholmen Science Park Aktiebolag, Mioeind EHF., Haskoli Islands, Norges Teknisk-Naturvitenskapelige Universitet (NTNU), og Nederlandse Organisatie Voor Toegepaste, Akademie für Künstliche Intelligenz.
3. Aktører Hvem udfører aktiviteterne? Hvilken afdeling af instituttet? Evt. hvilke eksterne parter er med (videninstitutioner, virksomheder, erhvervsorganisationer, myndigheder, klyngeorganisationer eller andre.)	Alexandra Instituttets medarbejdere fra to afdelinger vil udføre aktiviteterne: <i>Artificial Intelligence & Data Analytics Lab</i> og <i>Strategic Business & Governance</i>. Centrale eksterne samarbejdspartnere er: • DIKU, Datalogisk Institut ved Københavns Universitet • Digitaliseringsstyrelsen • AU, Aarhus Universitet, Center for Humanities Computing Aarhus • Konsortiet bag AI Denmark; Teknologisk Institut, Danmarks Tekniske Universitet, Aalborg Universitet, ITU, DIKU samt Innovation Centre Silicon Valley
4. Sammenhæng med andre projekter Indgår aktiviteten i andre eksternt finansierede projekter?	RK-indsatsen medfinansierer og sikrer sammenhæng til følgende igangværende projekter: • CoRAL, Grand Solution: Indsamling af taledata og udvikling af dansk tekst-til-tale model • TRUST-LLM, Horizon Europe: Udvikling af Germansk sprogmodel • AI-Matters, Digital Europe, TEF: Formålet er at udvikle en europæisk testplatform, hvor fremstillingsvirksomheder kan teste sine AI teknologier. • European Digital Innovation Hubs: TechCircle (CD-EDIH/Midtjylland), SEDIH (Syddanmark), AI Boost (GC-EDIH/Hovedstaden) Ovenstående projekter leverer 1) behovsafdækning, udvikling, modning og pilottest af relevante TDU services i RK-indsatsen som er under udvikling og 2) relevant vidensspredning til målgruppen.

	Øvrige projekter, der bl.a. bidrager med viden og adgang til RK indsatsens målgruppe: • AI Denmark, Industriens Fond: Formålet er at understøtte SMV'er med at komme hurtigere i gang med at udnytte data og AI-værktøjer i deres forretning. • Forskningsprojekt i samarbejde med Forsvaret (FMI) som blandt andet har fokus på anvendelsen af sprogmodeller
5. Følgegruppe <i>Har følgegruppen forholdt sig til aktiviteten? I så fald hvordan?</i>	Vi har haft løbende dialog med nøglespillere inden for det sprogteknologiske område i Danmark. Dette omfatter aktører som Omilon, Corti, KMD, ATP, Dansk Sprognævn, Digitaliseringsstyrelsen m.fl. Denne vedvarende dialog har været afgørende for at opretholde en omfattende forståelse af udviklingen inden for sprogteknologi og AI. Vi forventer at styrke vores samarbejde med aktørerne i løbet af 2024 på grund af synergier med blandt andet CoRal-projektet.
6. Formidling af resultater <i>Hvordan/hvor kan interesserede virksomheder m.fl. få viden om resultaterne af aktiviteterne?</i> <i>Anføres/tilføjes hvis det ikke allerede fremgår af beskrivelsen ovenfor, f.eks. ved links til konferencer, hjemmeside, publikationer etc.</i>	Resultater og erfaringer fra denne aktivitet vil primært blive formidlet gennem aktiviteten 'Demonstration af anvendelighed og værdiskabelse' inden for indsatsområdet Digital sikkerhed, tillid og dataetik. Formidlingen laves gennem forskellige formater såsom white papers, blogindlæg samt præsentationer på konferencer og webinarer. Desuden vil materialer være tilgængeligt gennem TDU-AI på alexandra.dk og Alexandras profil på Hugging Face. For at nå ud til en bredere målgruppe vil en betydelig del af formidlingen foregå i samarbejde med andre aktører og samarbejdspartnere.