Dansk AI

A. INDLEDENDE OPLYSNINGER	
Aktivitetsområde	Indsatsområdet Digital sikkerhed, tillid og dataetik
Institut	Alexandra Instituttet
Titel *Dækker indholdet af aktiviteterne*	Dansk AI
Nummerering *Af beskrivelsen*	3
Version	1
Periode *Forventet start og slut*	1/1 2023-31/12 2023
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 *Hvorfor? Hvad er målet for aktiviteterne? Hvordan bidrager de til det overordnede mål for indsatsområdet?*	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 Whisper-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 bidrages 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 vidanspartnere, samt vidanspredning af resultater til dansk erhvervsliv og andre interesserede i form af eksempelvis indlæg på konferencer, webinarer, formidlingsrapporter, artikler, blogindlæg, m.v.
2. Indhold **Hvad skal der ske?** **Hvilke(n) konkret(e) aktiviteter udføres?**	Med afsæt i ovenstående målbeskrivelse udføres følgende 3 delaktiviteter i 2023: **1. Udvikling af danske tekst og tale datasæt og modeller** Delaktiviteten har tre primære fokusområder. To af dem vil have et særligt fokus taledata og -teknologier. Det skyldes at arbejdet i 2022 har vist at området er både særligt udfordret og udfordringerne komplekse at løse for virksomhederne på egen hånd.[2] *(1.1) Produktion af danske taledata* Der har i en del år været efterspørgsel fra danske virksomheder på et større annoteret dansk taledatasæt.[3] Det skyldes dels at der i dag er få tilgængelige data, at det kan være svært at bruge egne data, fx grundet GDPR, samt at taledata er dyre at producere. I 2023 vil produktion af danske taledata derfor være et tema. Fokus vil være på bred repræsentation af talt dansk herunder minimering af uhensigtsmæssig bias mod fx dialekter og sprogstile som er en stor problematik i dag. Delaktiviteten overlapper med projektet CoRal (se nedenfor) og der vil desuden blive arbejdet på at skabe yderligere synergiprojekter, der kan geare den samlede mængde af taledata der produceres, da dette er en stor mangelvare hos virksomheder. *(1.2) Udvikling af danske talemodeller* I 2023 vil der blive udviklet forskellige talemodeller. Det kan fx være tale til tekst, tekst til tale, akustiske modeller eller domænespecifikke modeller, der kan være med til at eksemplificere brugen af taleteknologi indenfor forskellige industrier. I forhold til sidstnævnte modeltype vil de blive udviklet og afprøvet i et eller flere case-samarbejder med virksomheder og/eller offentlige organisationer. I forbindelse med udvikling af modellerne vil vi også eksperimentere med at udvikle såkaldte *destillerede* modeller (komprimerede versioner af ovenstående modeller, der i kraft af at være mindre kræver mindre computerkraft og energi at afvikle). Det kan være en stor fordel applikationsmæssigt og vil bidrage til at danske virksomheder hurtigere og billigere kan eksperimentere med forskellige taleteknologiske modeller i deres nuværende løsninger eller til at teste potentielle use cases af. *(1.3) Udvikling af danske grundmodeller* Grundmodeller er maskinlæringsmodeller, der er trænet på store mængder af ikke annoteret data. De bruges typisk som en base/grundmodel (deraf navnet) til træning af mere opgavespecifikke maskinlæringsmodeller, som kan være modeller der analyserer sentiment eller identificerer stødende sprogbrug, hvor grundmodellen finjusteres med et mindre og specifikt datasæt, der er repræsentativt for den pågældende opgave. Disse grundmodeller har afgørende indflydelse på den kvalitet, der kan opnås med de opgavespecifikke modeller. Det gælder både i forhold til hvor præcis modellen er, men også i forhold til at modvirke uhensigtsmæssige bias. En af de mest benyttede åbne danske grundmodeller er fra 2018 og der er derfor behov for at træne nye modeller, der baserer sig på de seneste modelarkitekturer og metoder. Da træning af grundmodeller er omkostningstung, både i computerkraft, energiforbrug og klimapåvirkning, er det oplagt at gøre sådanne modellen åben tilgængelige, så de benyttes af alle.

[2] https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_9__Language_Report_Danish_.pdf
[3] https://dsn.dk/wp-content/uploads/2021/01/sprogteknologi-i-verdensklasse.pdf

	(2) Udvikling og lancering af TDU for dansk data science og integrering af DaNLP I 2022 er der udviklet prototype på en TDU (Test, Demonstration og Udvikling) for dansk data science (for nuværende kaldet AIAI (Alexandra Instituttet Artificial Intelligence)). TDU'en har fokus på Dansk AI generelt, men i første omgang tekst- og taledata. Denne delaktivitet har til formål at løfte den nuværende TDU fra prototype til en første driftsklar version. TDU'en er målrettet machine learning-udviklere i danske virksomheder, som enten allerede arbejder med dansk data science eller som gerne vil i gang. DaNLP integreres desuden i TDU'en. TDU'en leverer end-to-end rådgivning og udvikling af softwareløsninger baseret på tekst- og taledata indenfor sprogteknologi med særligt fokus på dansk. Dertil indeholder TDU'en en række tekniske funktioner, som understøtter danske virksomheder i (1) at skabe overblik over tilgængelige open source-muligheder, så de kan træffe det bedste tekniske valg for dem og (2) afhjælpe og minimere de udfordringer, som udviklere ofte støder på. Målet er at lette både udvikling, opdatering og vedligehold af maskinlæringsbaserede systemer (som i nogle tilfælde kan være forskellen på en positiv og en negativ business case). De første og allerede tilgængelige funktioner i AIAI giver mulighed for på en let, intuitiv og hurtig måde: 1. At identificere de bedst egnede danske modeller indenfor specifikke opgaver. 2. At evaluere egne eller eksisterende modeller på parametre som fx præcision, hastighed, energiforbrug og klimaaftryk. Yderligere funktioner der overvejes er: 3. Services der gør det muligt at udstille modeller lokalt og/eller i skyen. 4. Præfabrikerede pipelines, så basal kode (boilerplate) ikke skal skrives igen og igen af forskellige udviklere. 5. Produktion af åbne danske datasæt. 6. Kuratering af åbne danske datasæt så de er lettere at tage direkte i brug. 7. Kontinuerlig opdatering af de bedste modeller på dansk indenfor en forskellige NLP-/taleteknologiske opgaver. 8. Compliance værktøjer, som for eksempel anonymisering af dokumenter. De forskellige funktioner vurderes og ændres løbende og kvalificeres med både markedsaktører, den kommende følgegruppe og andre interessenter, så der ikke udvikles funktioner, der er markedsforvridende. **(3) Udvikling af samarbejde med vidanspartnere** Samarbejde med både DIKU og AU fra de foregående år forsættes og udbygges i 2023. Begge har særligt fokus på udvikling af taledata og talemodeller. Der vil blive afsat en mindre del af budgettet for aktiviteten til udvikling af synergiprojekter og ansøgningsskrivning.
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: *Artificial Intelligence & Data Analytics Lab* og *Strategic Business & Governance.* 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 Sillicon Valley
4. Sammenhæng med andre projekter *Indgår aktiviteten i andre eksternt finansierede projekter?*	Der er synergi med følgende projekter: • **Co**nversational and **R**ead-**al**oud speech dataset (CoRal), der er støttet af Innovationsfonden og har til formål at producere åbne danske taledata i høj kvalitet

	• AI Denmark projektet hos Industriens Fond, der har til formål at understøtte SMV'er med at komme hurtigere i gang med at udnytte data og AI-værktøjer i deres forretning.
5. Følgegruppe *Har følgegruppen forholdt sig til aktiviteten? I så fald hvordan?*	I 2023, Q1 etableres en følgegruppe specifikt for denne aktivitet. I løbet af 2022 er indholdet af aktiviteten blevet forholdt og diskuteret med væsentlige aktører indenfor det sprogteknologiske område i Danmark såsom Omilon, Corti, Dictus, Alvenir, KMD, ATP, Dansk Sprognævn, Digitaliseringsstyrelsen, m.fl.
6. Formidling af resultater *Hvordan/hvor kan interesserede virksomheder m.fl. få viden om resultaterne af aktiviteterne?* *Anføres/tilføjes hvis det ikke allerede fremgår af beskrivelsen ovenfor, f.eks. ved links til konferencer, hjemmeside, publikationer etc.*	Resultater fra aktiviteten udstilles via Alexandras profil på Hugging Face. Derudover indtænkes resultaterne fra aktiviteten i Instituttets generelle vidensformidling og i arbejdet omkring den videre etablering af TDU'en.