DIGITALISERINGSSTYRELSEN

Overvejelser til opgavebeskrivelse

29. oktober 2020

Overvejelser til opgavebeskrivelse til et dansksproget transskriberet og tidskodet talekorpus

1. Indledning

Der er stor efterspørgsel i de sprogteknologiske miljøer efter talegenkendelsesressourcer og sprogdata af høj kvalitet, der kan anvendes til træning af algoritmer i forbindelse med udvikling af dansksprogede talegenkendelsessystemer, der transformerer tale til tekst. Sprogteknologiudvalget anbefaler i forlængelse heraf, at der offentligt investeres i udvikling og udstilling af sprogressourcer af høj lingvistisk kvalitet. Udvalget peger særligt på behovet for udvikling af en generel, tidskodet og kvalitetssikret taleressource, der kan danne basis for, at leverandører kan udvikle mere specifikke løsninger inden for dansk talegenkendelse.[1]

Udviklingsprojektets formål er at udvikle en dansksproget talegenkendelsesressource i form af et stort transskriberet og tidskodet tekst- og talesprogskorpus, der kan være med til at løfte niveauet for dansk talegenkendelse.

Nærværende dokument udstikker rammerne og processen for udviklingsprojektet herunder beskriver en række valg, der er foretaget, mens andre valg skal specificeres nærmere i samspil med interessenter og fagpersoner. Udviklingsprojektet indebærer, at opgaven med at indsamle lyddata samt bearbejdning af tale- og tekstkorpora skal sendes i udbud.

Målgruppen for anvendelse af talekorpus er primært sprogteknologiske virksomheder samt offentlige myndigheder og private virksomheder, der udvikler eller ønsker at udvikle eller forbedre dansksprogede talegenkendelsesløsninger. Digitaliseringsstyrelsen bliver ejer af det færdigudviklede tekst- og talesprogskorpus, der vil blive stillet frit tilgængeligt for alle via sprogteknologi.dk.

2. Baggrund

I regi af Digitaliseringspagten og Økonomiaftalen 2020 blev det aftalt at igangsætte og finansiere et fællesoffentligt samarbejde om udviklingen af en fælles

[1] Sprogteknologiudvalget blev nedsat af Dansk Sprognævn under Kulturministeriet og udgav i april 2019 rapporten 'Dansk sprogteknologi i Verdensklasse'.

dansk sprogressource, der i dag bærer navnet sprogteknologi.dk[2]. Formålet med samarbejdet er at understøtte udviklingen af dansksproget kunstig intelligens.

Formålet med sprogteknologi.dk er at gøre relevante, eksisterende sprogressourcer tilgængelige for alle digitalt samt udvikle og udstille nye sprogressourcer, som kan styrke udviklingen af sprogteknologiske løsninger på dansk. Sprogressourcerne skal kunne tilgås digitalt, let og fleksibelt og således understøtte udviklingen af innovative dansksprogede løsninger i samspil med markedet.

Udviklingen af sprogteknologi.dk tager bl.a. udgangspunkt i de anbefalinger, som Sprogteknologiudvalget kom med i rapporten 'Dansk Sprogteknologi i Verdensklasse' (april 2019). Udvalget anbefaler en række initiativer, bl.a. et tidskodet dansk talesprogskorpus, som vil gøre det muligt at træne sprogteknologiske løsninger til intelligent taleforståelse, som for eksempel talegenkendelse, talesyntese, taleridentifikation, IVR (interactive voice-response) og dialogsystemer.[3]

3. Afgrænsning

Overordnet er udviklingsprojektet afgrænset til at skulle understøtte udviklingen af dansksproget talegenkendelse. Desuden er indsamlingsmetoden i første omgang afgrænset til 'generelt sprog' (fase 1) med muligheden for, at der på sigt kan bygges ovenpå (evt. fase 2) med mere domænespecifikke sprog, fx målrettet kommunal sagsbehandling, sundhed, jura, etc.

Derudover angives i afsnit 4 en række krav, der yderligere afgrænser udviklingsprojektet.

4. Overvejelser om opgavebeskrivelse

Talegenkendelsesressourcen skal kunne anvendes som datagrundlag for forskellige typer af maskinlæring. Dette gælder både dyb læring (supervised learning), som kræver data af høj kvalitet, der er opmærket med metadata af høj kvalitet, og overfladeorienteret læring (unsupervised learning) som kræver meget store datamængder, samt kombinationer af de to metoder.

Desuden skal ressourcen kunne understøtte udvikling af talegenkendelsesløsninger målrettet forskellige brugssituationer, herunder især den professionelle anvendelse og den brede almene anvendelse. Den professionelle anvendelse er kendetegnet ved lav fejltolerance, men bruger udstyr af god kvalitet og arbejder med begrænset baggrundsstøj. For at opnå det bedste resultater i denne anvendelsessituation skal det lydmateriale, der anvendes til træning, være optaget på udstyr af god kvalitet med minimal baggrundsstøj. Den brede anvendelse er kendetegnet ved, at

[2] https://sprogteknologi.dk/
[3] Dansk Sprognævn, 'Dansk Sprogteknologi i Verdensklasse' (april 2019), s. 68-69.

kvaliteten af det udstyr, der tales i, er meget varieret, og mængden af baggrundsstøj kan ligeledes være meget varieret. Til gengæld er fejltolerancen typisk lidt højere. Til dette formål kan det være en fordel, at der ligeledes er variation i lydmaterialet.

Fremstillingen af et sådant talesprogskorpus har en række faser, hvoraf indsamling af lyddata er én fase. Den klart mest omkostningstunge del følger efter indsamlingen af lyddata, idet et talesprogskorpus – for at kunne have værdi som sprogteknologisk komponent – skal annoteres med lydskrift og tidskodes af fonetiske eksperter.

4.1 Metode
Med afsæt i ovenstående forventes det, at leverancen skal bestå af to dele:

Del 1: Det er forventningen, at der skal anvendes minimum 200 timers tale-korpus i høj lydkvalitet med meget høj kvalitet af transskription og tids-kodning, som primært evalueres på kvaliteten.

Del 2: Det er forventningen, at der skal anvendes minimum 2.000 timers tale-korpus, hvor en mindre høj kvalitet er acceptabel, og hvor det vurderes positivt, hvis leverandøren leverer yderligere timer.

Det er hensigten, at leverandøren vil få metodefrihed til, hvordan lydmaterialet skal indsamles, fx om de vil anvende allerede indsamlet materiale, crowdsourcing, udvalgte informanter, etc. Dog skal de færdige korpora overholde visse krav, herunder skal de:

- være repræsentative, dvs. indeholde talere af forskellig køn, aldre, geografiske og sociale baggrunde, samt talere med talefejl og som taler dansk som fremmedsprog
- dække samtlige danske sproglyde med en variation som beskrevet ovenfor
- indeholde både planlagt tale (oplæsning) og spontantale
- være fri af juridiske bindinger, så data kan anvendes helt frit både kommercielt og til forskning
- angive metadata for de enkelte optagelser om speaker og optagelsessituation
- afleveres på en velorganiseret og veldokumenteret måde, der giver mulighed for tilføjelse af yderligere data
- afleveres i et velegnet og åbent format, fx .wav for lydfilerne og .txt for transskriptionerne

I den videre proces skal disse dette præciseres nærmere, og der skal desuden tages stilling til, om der skal sættes krav til:

- det emnemæssige og teksttypologiske indhold
- inklusion af dialog i den ene eller begge dele af ressourcen
- pligtindhold i form at tekster, der er designet til at dække alle danske sproglyde, og som ressourcen skal indeholde et vist antal oplæsninger af
- minimum antal talere og/eller et vist minimumsbidrag per taler
- inklusion af supportværktøjer i afleveringen
- kompetencer hos leverandøren, fx at det skal være personer tilknytte med fonetisk ekspertise, dansk på indfødt niveau og viden om dokumentation og organisering af data