DIGITALISERINGSSTYRELSEN

Spørgeramme til markedsdialog

21. oktober 2020

Nedenstående spørgeramme vil danne udgangspunkt for en dialog på mødet. Det er således ikke en forudsætning, at alle spørgsmål skal besvares, idet hensigten med markedsdialogen er, at Digitaliseringsstyrelsen får realistiske og velfunderede input til udbuddet, herunder fx hvilke behov markedet ser i forbindelse med udviklingen af et dansksproget transskriberet og tidskodet talekorpus, hvilke særlige hensyn, der bør overvejes, hvad der bør prioriteres samt hvilke teknologiske muligheder, der eksisterer og er realistiske inden for den givne økonomiske ramme.

1.1 Krav til slutproduktet - et dansksproget transskriberet og tidskodet talekorpus

1.1.1 Kunne jeres virksomhed tænkes at benytte korpus og til hvilke formål?

1.1.2 Har I forslag til krav til omfang, kvalitet (lydkvalitet, transskriptionskvalitet) og sammensætning, der skal være opfyldt for, at talekorpusset vil give værdi for jer? Har I nogle kommentarer til hvilke krav, der er vigtigst at få opfyldt?

1.1.3 Det er hensigten, at datasættet skal repræsentere et bredt udsnit af den danske befolkning (køn, alder, dialekter, etc.), så det kan bruges til at træne talegenkendelse, der kan anvendes af alle. Er der nogle befolkningsgrupper eller kriterier, vi i den forbindelse bør være særligt opmærksomme på?

1.1.4 Digitaliseringsstyrelsen påtænker, at talekorpusset skal bestå af to dele. En del i meget høj lydkvalitet med perfekt ortografisk og fonetisk transskription egnet til superviseret læring og forskning, og en større del med mindre rigide kvalitetskrav og blot ortografisk transskription, der egner sig til usuperviseret læring. Har I nogen kommentarer til dette?

1.1.5 Er der nogle særlige fonetiske fænomener, vi bør være opmærksomme på (ud over dækning af danske fonemer og "lovlige" fonemkombinationer)?

1.1.6 Har I andre krav eller ønsker til sammensætning af data? Fx forholdet mellem planlagt og spontan tale/monolog og dialog, eller indeholdte teksttyper.

1.1.7 Har I krav eller ønsker til omfanget at metadatadata? Fx data om informanter, optagesituationer eller -udstyr.

1.1.8 Har I andre krav eller ønsker til udformningen af data? Fx til anvendte formater, optageudstyr eller standarder for transskription og metadata?

1.2 Opgavetager

1.2.1 Kunne jeres virksomhed være interesseret i at byde på opgaven? Har I nogle kommentarer i forhold til, hvad der kunne få jer til - eller forhindre jer i - at byde på den?

1.2.2 Digitaliseringsstyrelsen påtænker i forbindelse med tilbuddet at udbede en mindre sampletest af transskribering, annotering og lyd, som et eksempel på den kvalitet, tilbudsgiver kan levere. Vil det være en god idé eller risikere at afskrække en potentiel tilbudsgiver? Er der andre metoder til at sikre, at tilbudsgiver er i stand til at levere den fornødne kvalitet?

1.2.3 Hvad er rimeligt at bede om i form af referencer til tidligere udførte sammenlignelige opgaver? Hvad skal tælle som sammenlignelige opgaver, og hvor mange referencer kan vi bede om, når formålet er at udelukke tilbudsgivere uden forståelse for opgaven, men ikke at udelukke små eller nystartede virksomheder?

1.3 Udarbejdelse af korpus

1.3.1 Hvilke metoder forstiller I jer der kan anvendes til udvikling af et talekorpus? Herunder sammensætning af metoder? (fx crowdsourcing, anvende eksisterende data m.v.)

1.3.2 Digitaliseringsstyrelsen forestiller sig, at en løsning som den beskrevne, ville kunne anskaffes til en pris mellem DKK 2 og 2,5 mio., ekskl. moms, alt inkl. Hvilke væsentlige cost-drivere mener I, der særligt er i forbindelse med udarbejdelsen af det beskrevne korpus?

1.3.3 Digitaliseringsstyrelsen forestiller sig, at der afsættes ca. 1½ år til, at leverandøren kan udvikle den endelige løsning. Er det en rimelig periode for det beskrevne?

1.3.4 Digitaliseringsstyrelsen forestiller sig, at der undervejs i forløbet afleveres delleverancer, dels så Digitaliseringsstyrelsen kan sikre, at der er den fornødne fremdrift og kvalitet i udviklingen, dels med forhåbning om, at dele af korpus kan udstilles og nyttiggøres tidligere. Har I nogen kommentarer til, hvordan delleverancer bør struktureres?

1.3.5 Har I forslag til, hvordan korpus bedst kvalitetssikres?

1.3.6 Digitaliseringsstyrelsen ønsker at sætte mindstekrav til opgaveløsningen således, at korpus samt alle relaterede værktøjer:

- ejes af Digitaliseringsstyrelsen ved opgaveløsningens ophør
- må anvendes og udstilles frit
- er udviklet i frie formater og open source

Har I nogen kommentarer til det?

1.4 Andet

1.4.1 Har I kommentarer til emner, der er relevante for udbuddet af et dansksproget transskriberet og tidskodet talekorpus, der ikke er dækket af de øvrige spørgsmål?

1.4.2 Ser I nogle særlige risici i forbindelse med udviklingen af et dansksproget talekorpus?

1.4.3 Hvordan undgås det, at stride mod GDPR i forbindelse med tilvejebringelse af (lyd)data?