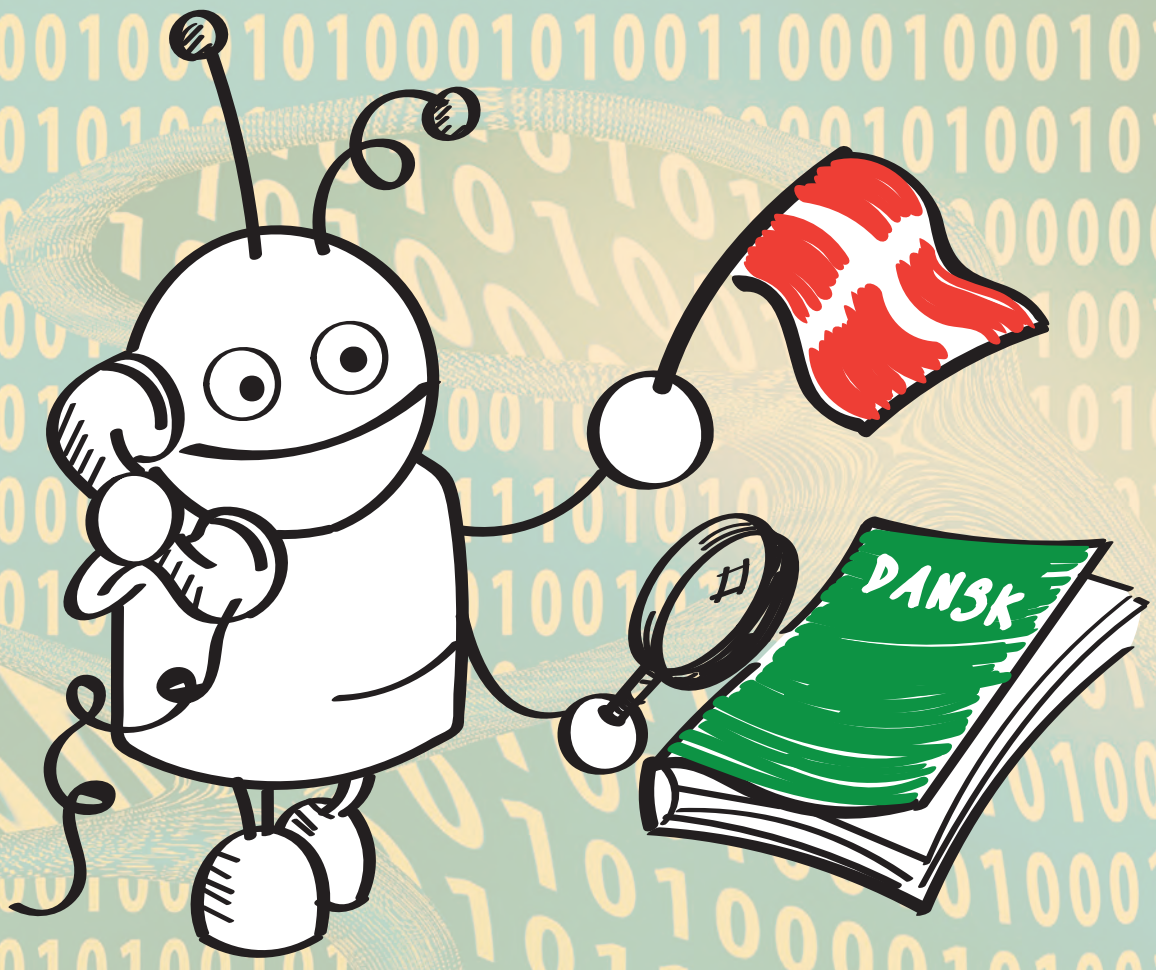
DANSK SPROGTEKNOLOGI I VERDENSKLASSE



Dansk Sprognævn

SPROGTEK2018

Sprogteknologiudvalget under Dansk Sprognævn

nedsat af Kulturministeriet

Dansk Sprogteknologi i Verdensklasse

Rapport fra sprogteknologiudvalget under Dansk Sprognævn nedsat af Kulturministeriet

Denne rapport er resultatet af arbejdet i det sprogteknologiske udvalg under Dansk Sprognævn; udvalget blev nedsat i slutningen af 2017 af kulturminister Mette Bock.
Dansk Sprognævn har fungeret som udvalgets sekretariat.
Sprognævnets direktør Sabine Kirchmeier har været formand for udvalget.

Rapporten er forfattet af Sabine Kirchmeier, Peter Juel Henrichsen, Philip Diderichsen og Nanna Bøgebjerg Hansen.
April 2019.

Dansk Sprogteknologi
i Verdensklasse
Rapport fra sprogteknologiudvalget
under Dansk Sprognævn
nedsat af Kulturministeriet

ISBN 978-87-89410-77-7

Dansk
Sprognævn

Indhold

Bevarelsen af et sprog og dermed af den kultur, der udvikler sig omkring det, er i høj grad betinget af dets evne til at fungere og være nyttigt i moderne og foranderlige miljøer som den digitale verden.

Fra UDKAST TIL BETÆNKNING om ligebehandling af sprog i en digital tidsalder (2018/2028(INI)). EU-Parlamentet.

Sammenfatning

Sprogteknologiudvalget under Dansk Sprognævn blev nedsat af kulturministeren i slutningen af 2017. Formålet var at kortlægge behov og muligheder for at bruge dansk i teknologier som anvender sprog og kunstig intelligens. Udvalget skulle endvidere afklare perspektiverne for en dansk termbank.

Den danske regering præsenterede i efteråret 2018 sine visioner for digital service i verdensklasse som bl.a. indebærer at Danmark skal gå forrest i anvendelsen af kunstig intelligens herunder dansk sprogteknologi. Konkret satses der på opbygningen af en sprogresurse på dansk der sættes til fri afbenyttelse så leverandørerne har en fælles sprogresurse af høj kvalitet der giver dem mulighed for at udvikle gode løsninger inden for talegenkendelse og sprogforståelse med et højt præcisionsniveau.

Udvalgets anbefalinger forholder sig til begge initiativer.

En stor del af vores viden er formuleret på et sprog. Størstedelen af den viden vi har om Danmark, om danske forhold og om hinanden, er formuleret på dansk. Kunstig intelligens er typisk baseret på analyse af store datamængder. Det giver gode resultater når disse data er tal, men det er en langt større udfordring når data består af sprog i form af tekst og lyd. Tal er entydige og svarer til den måde computerne er indrettet på. Sprog er mangetydigt og langt mere komplekst fordi det er en del af vores eksistens og tæt sammenvævet med vores viden om verden, om den måde vores samfund er opbygget på, og om den kultur vi er opvokset i.

Mange systemer i Danmark som involverer sprog og kunstig intelligens, er enten udviklet på basis af data fra andre sprog, først og fremmest engelsk, eller på basis af danske data af lav kvalitet. Hvis udgangspunktet ikke er danske data, risikerer man at systemerne reflekterer en forestilling om en verden, et samfund og en kultur som ikke stemmer overens med vores. Hvis udgangspunktet er danske data af lav kvalitet, risikerer man at kvaliteten bliver så ringe at brugerne afviser den nye teknologi. Derfor skal systemer som involverer sprog og kunstig intelligens, udvikles på basis af danske sprogdata af høj kvalitet for at kunne fungere optimalt og gøre nytte i vores samfund.

Dansk adskiller sig markant fra engelsk både i den måde vi danner ord og sætninger på, og i den måde vi udtaler ordene på. Dansk er rigere på vokaler end mange andre sprog. Vores udtale er i høj grad præget af fonetiske reduktioner – det som udlændinge ofte oplever som mumleri. Det gør det danske sprog vanskeligt at arbejde med - især for taleteknologien. Det er derfor nødvendigt at dansk sprogteknologi udvikles af mennesker som behersker dansk.

Der har igennem de sidste årtier været afsat en del midler til forskning i dansk sprogteknologi. Der er udviklet en række sprogresurser af høj kvalitet med offentlig støtte, og der findes en række store og små danske og internationale virksomheder som producerer dansk sprogteknologi til det danske marked.

Men initiativerne er spredte og ukoordinerede, og sprogresurserne er kun begrænset tilgængelige.

Med kun ca. 5,6 mio. dansktalende i verden, de fleste af dem i Danmark, udgør det danske sprog og dermed Danmark et meget lille marked både for store internationale sprogteknologivirksomheder og for danske virksomheder. Der skal derfor en særlig indsats til for at sikre at der udvikles sprogteknologi i høj kvalitet for det danske sprog.

Det er et helt grundlæggende vilkår for al sproglig udvikling at sproget bliver brugt i alle samfundets områder. Hvis sproget ikke bruges, mister det sin evne til at udtrykke det som vi har behov for, på en smidig og hensigtsmæssig måde.

Det danske sprog i sig selv er ikke umiddelbart truet hvis vi ikke får adgang til god sprogteknologi på dansk. Men der vil være et stigende antal sammenhænge hvor danskerne enten skal stille sig tilfredse med dårlig og fejlagtig sprogbrug, misforståelser og fejl fra systemernes side, eller de skal skifte over til at bruge engelsk. Sammen med den stigende brug af engelsk i andre dele af samfundet, fx på universiteter og i erhvervslivet, kan dette på længere sigt føre til en ringere funktionsdygtighed af det danske sprog.

Udvalget ser imidlertid de største effekter og gevinster i forhold til de mennesker der bruger sproget, frem for i forhold til selve sproget. Den største risiko ligger i de samfundsmæssige konsekvenser og i konsekvenserne for den enkelte hvis der ikke udvikles sprogteknologi af god kvalitet for dansk.

Sprogteknologiudvalget anbefaler derfor:

- At der oprettes en organisation som har til opgave at koordinere indsatsen for dansk sprogteknologi
- At der oprettes en dansk sprogbank som skal understøtte udviklingen og vedligeholdelse af danske produkter baseret på sprogteknologi og kunstig intelligens, bl.a. ved at gøre danske sprogresurser og sprogværktøjer frit tilgængelige, herunder en dansk termbank
- At uddannelse i dansk sprogteknologi prioriteres
- At forskning i dansk sprogteknologi styrkes.

Udviklingen inden for kunstig intelligens og sprogteknologi går hurtigt, og der tages løbende nye algoritmer og metoder i brug. Udvalget har derfor lagt vægt på fleksible løsninger der kan opdateres løbende med stadigt fokus på løsningernes samfundsmæssige effekt.

1. Sprogteknologiudvalgets arbejde

Sprogteknologiudvalget blev nedsat af kulturministeren under Dansk Sprognævn i slutningen af 2017 og påbegyndte sit arbejde den 1.1.2018. Udvalgets arbejde er finansieret som en del af en bevilling på Finansloven på 0,5 mio. kr. til styrkelse af dansk sprogteknologi.

1.1. Sprogteknologiudvalgets kommissorium

Udvalget havde til formål at udrede perspektiver og udfordringer for sprogteknologi i en dansk kontekst og komme med forslag til hvordan Danmark bedst sikrer brugen af dansk og andre sprog i digitale tjenester, fx automatisk oversættelse, taleteknologi, IoT (internet of things), robot- og transportmiddelteknologi, it-baserede læremidler til sprogundervisning og kunstig intelligens.

Udvalget havde endvidere til formål at afklare behovet og perspektiverne for en national termbank ("sprogtermbank").

Udvalget skulle inddrage relevante resultater fra arbejdet med sprogteknologi og terminologi i andre lande, herunder EU og Norden, og pege på måder hvorpå en styrkelse af dansk sprogteknologi vil kunne gavne den enkelte borger og bidrage til at skabe vækst og effektivisering i samfundet.

Udvalget fik til opgave at

1. levere en rapport der udreder behovet for sprogteknologi inden for centrale sektorer. Rapporten skulle give svar på følgende hovedspørgsmål:
   - Inden for hvilke sektorer og erhverv vil der i de kommende 10 år være størst behov for digitale tjenester og applikationer baseret på kunstig intelligens på dansk og andre sprog?
   - Hvilke udfordringer ser virksomheder og offentlige institutioner i forhold til at udvikle disse tjenester og applikationer – og hvilke udfordringer bliver overset?
   - På hvilken måde kan sprogteknologi bidrage til at sikre en bedre og billigere offentlig service?
   - På hvilken måde kan erfaringer fra andre lande, EU og Norden nyttiggøres?
   - Hvilke vækst- og jobmuligheder ligger der i en satsning på dansk sprogteknologi?
   - Hvad er den samfundsøkonomiske business case set i forhold til investeringsbehovet?
   - Hvilke politiske tiltag kan foreslås for at understøtte virksomheder og offentlige institutioner i at inddrage dansk og andre sprog når der skal udvikles og anvendes nye teknologier baseret på kunstig intelligens?
   - Hvilken betydning får en satsning på dansk sprogteknologi for udviklingen af det danske sprog, for samfundets udvikling og for den enkelte?
   - Hvordan sikres det at der udvikles dansksproget sprogteknologi?
   - Hvad er fordelene og ulemperne ved udvikling af dansk sprogteknologi i Danmark?
   - Hvordan kan det sikres at der fortsat uddannes mennesker med tilstrækkelige kompetencer inden for dansk sprogteknologi?
   - Hvilket behov er der for udvikling af en dansk termbank, hvilke domæner skal den dække, og hvordan kan den bedst gøres tilgængelig?
2. bidrage til oplysning og offentlig debat om sprogets rolle i kunstig intelligens og ny teknologi
3. inddrage offentlige institutioner, virksomheder, brancheforeninger, fagforeninger, fageksperter og borgere med henblik på at sikre at så mange aspekter som muligt bliver belyst.

1.2. Udvalgets sammensætning

Udvalget blev nedsat som en bredt sammensat arbejdsgruppe bestående af 14 medlemmer inkl. formand og en sekretær. Sekretariatsfunktionen og formandskabet blev forankret i Dansk Sprognævn.

Blandt medlemmerne var repræsentanter for nuværende og fremtidige udbydere og brugere af applikationer baseret på sprogteknologi og kunstig intelligens i erhvervslivet og den offentlige sektor, repræsentanter for udviklere af sprogteknologi og kunstig intelligens samt repræsentanter for forsknings- og uddannelsessektoren.

1. CTO Klaus Akselsen, MIRSK
2. Forskningsleder Esben Alfort, Ankiro ApS
3. Udviklingschef Lars Fremerey, GTS-foreningen
4. Computational Linguist Anna Katrine Jørgensen, Google
5. Sekretariatschef Jens Kellerup, Ballerup Kommune/OS2 – (Offentligt digitaliseringsfællesskab)
6. Direktør Sabine Kirchmeier, Dansk Sprognævn (formand for udvalget)
7. Direktør Jens Otto Kjærum, Dictus
8. Professor Bodil Nistrup Madsen, CBS – Copenhagen Business School
9. Seniorredaktør Sanni Nimb, Det Danske Sprog- og Litteraturselskab
10. Professor Bolette Sandford Petersen, Center for Sprogteknologi, Københavns Universitet
11. Forsknings- og Innovationsdirektør Anders Quitzau, IBM Research – Watson Advocate
12. Founder, Chief Visionary Officer Mads Rydahl, Unsilo
13. Kontorchef Jens Krieger Røyen, Digitaliseringsstyrelsen (siden afløst af Anders Munk Andersen)
14. Chefkonsulent Carl Østergaard, Odense Kommune

Udvalgssekretariatet bestod af Sprognævnets direktør, Sabine Kirchmeier, seniorprojektforsker Peter Juel Henrichsen, videnskabelig it-medarbejder Phillip Diderichsen og studentermedhjælp Nanna Bøgebjerg Hansen.

1.3. Udvalgets arbejdsmetode

Udvalget har sigtet mod at afdække den aktuelle situation for dansk sprogteknologi fra 4 forskellige perspektiver: brugere, leverandører/producenter af sprogteknologi, udviklere af sprogteknologi samt forskere og undervisere i sprogteknologi. I alt har repræsentanter for ca. 120 forskellige danske virksomheder, organisationer, offentlige myndigheder og uddannelsesinstitutioner bidraget til udvalgets udredning.

Udvalgets vigtigste instrument til videnindsamling har været en række workshops med fokus på de 4 perspektiver suppleret med 2 workshops om specifikke områder, nemlig maskinoversættelse og terminologi. Som optakt til hver workshop har udvalget udsendt et spørgeskema. Data fra besvarelserne har givet grundlag for de kvantitative aspekter af sprogteknologiudvalgets udredning. Gennem strukturerede diskussionsoplæg har workshopdeltagerne afdækket de konkrete problemområder inden for hvert interesseområde samt foreslået de mest effektive tiltag.

Derudover har sekretariatet foretaget strukturerede interviews med relevante nøglepersoner i ind- og udland og inddraget rapporter og aktuelle informationer om arbejdet med sprogteknologi i Finland, Island, Letland, Norge og Sverige samt om de initiativer der foregår i EU-regi.

Endvidere blev der i januar 2019 afholdt et afsluttende seminar hvor alle workshopdeltagerne fik mulighed for at diskutere de anbefalinger som udvalget var nået frem til.

Resultatet af spørgeskemaundersøgelserne og workshopperne har sammen med interviews med relevante aktører og en grundig gennemgang af artikler og andre dokumenter om sprogteknologi dannet grundlaget for udvalgets rapport.

Udvalgets arbejde blev løbende formidlet via bloggen sprogtek2018.dk, og der blev produceret en kort tegnefilm om sprogteknologi som blev udsendt på bloggen, på Sprognævnets hjemmesider og via de sociale medier Facebook og LinkedIn.

Udvalget har holdt tæt kontakt med nordiske og europæiske organisationer der arbejder med sprogteknologi, for at dele viden på tværs af grænserne og afsøge samarbejdsmuligheder.

1.4. Konklusioner på udvalgets workshops

Deltagerne på udvalgets 6 workshops har i stort omfang bidraget til at tegne et aktuelt billede af situationen for dansk sprogteknologi.

Slutbrugerne, især organisationer i den offentlige sektor, oplever at de har ringe indflydelse på design og implementering af de sprogteknologiske grundprodukter (fx til talegenkendelse og oversættelse) som ofte udvikles af internationale firmaer. Mange tilkendegiver at de bidrager med sproglige data til udviklingsarbejdet, men at de ikke har råderet over deres data når de først er indarbejdet i et produkt. Dette medfører en stor afhængighed af udbyderne og begrænser slutbrugernes mulighed for at konkurrenceudsætte ydelsen betydeligt. Slutbrugerne melder ligeledes om problemer med systemernes kvalitet og som følge deraf vanskeligheder ved at få medarbejderne til at tage produkterne effektivt i brug.

Leverandørerne, både danske og udenlandske virksomheder, beskriver markedet som småt, men velorganiseret (især den offentlige sektor) og med nærhed til slutbrugerne. Det er en styrke at kunder i stat og kommune ofte selv deltager aktivt i udviklingsarbejde og finansiering. Blandt truslerne nævnes især at små virksomheder ofte fravælges som leverandører ved store udbud, hvilket især rammer de virksomheder som arbejder med dansk som deres hovedområde. Manglen på kvalificeret personale (danske sprogteknologer) bliver stadigt tydeligere. Det største problem er at alle virksomheder uanset størrelse skal ofre store resurser på at indsamle og udvikle de sproglige basisresurser for dansk som danner grundlag for produkterne. Mange peger på at det vil fremme udvikling af nye produkter og højne kvaliteten betydeligt hvis et sæt basisresurser af høj kvalitet var frit tilgængeligt og løbende blev vedligeholdt.

Udviklerne, ansatte i danske og udenlandske virksomheder og i danske forskningsinstitutioner, efterlyser fri adgang til danske basisresurser, herunder tekst- og talematerialer fri for bindinger (copyright, personbeskyttelse), samt strukturerede ordbaser med det danske ordforråds betydninger og udtaler. Udviklerne foreslår en uafhængig rådgivnings- og serviceorganisation der løbende kan udvikle, distribuere og vedligeholde danske sprogresurser og formidle viden om og levere kurser og vejledning i metoder og teknologier til håndtering af det danske sprog.

Forskerne peger især på at der mangler sprogteknologer med ekspertviden om det danske sprog og dermed på behovet for flere uddannelser og kursustilbud. Der peges på nødvendigheden af en organisation som kan planlægge udviklingen af nye basisresurser, kompetenceopbygning og formidling af sprogteknologi til det danske samfund.

Brugere af oversættelsesprogrammer understreger at det er nødvendigt især for offentlige institutioner at få en større bevidsthed om den værdi som deres sproglige data, fx oversatte tekster og terminologidatabaser eller -lister, repræsenterer for udvikling af sprogteknologi på dansk. Data som kan deles, bør identificeres og deles frit både til sprogteknologiske applikationer og applikationer som på anden måde anvender kunstig intelligens. Offentlige institutioner ses som helt afgørende spillere for fx at forbedre EU's oversættelsesprogram som stilles gratis til rådighed for alle offentlige institutioner i Europa, men flere aktører i den offentlige sektor har i den seneste tid outsourcet deres oversættelsesopgaver uden at sikre et tilbageløb af de strukturerede datasæt som produceres som et led i oversættelsesprocessen.

Brugere af dansk terminologi i den offentlige og private sektor peger enstemmigt på behovet for at terminologiarbejdet i Danmark bliver bedre koordineret for at undgå resursespild, og at det bedst kan gøres ved at oprette en national termbank. Størstedelen af terminologiarbejdet foregår i forbindelse med kommunikations- og oversættelsesopgaver, men udvalgets undersøgelser har vist at ca. 20 % af terminologiarbejdet foregår i forbindelse med udvikling af it-systemer og digitaliseringsklar lovgivning. Der er ikke blot behov for termer, definitioner og oversættelser, men der meldes i høj grad også om behov for struktureret viden om be-

grebernes relationer til hinanden. Blandt de fagområder som hyppigst nævnes som dem der er størst behov for, er jura, digitalisering, offentlig forvaltning, økonomi og sundhed.

1.5. Danmark i sammenligning med andre lande

Sprognævnets undersøgelser af udviklingen på sprogteknologiområdet i sammenligning med andre lande har givet følgende resultater: Det tekniske og resursemæssige grundlag for udvikling af dansk sprogteknologi er ikke på højde med forholdene i de sprogsamfund vi ofte sammenligner os med (fx Norge, Sverige, Finland og Nederlandene). Dette skyldes især fem forhold:

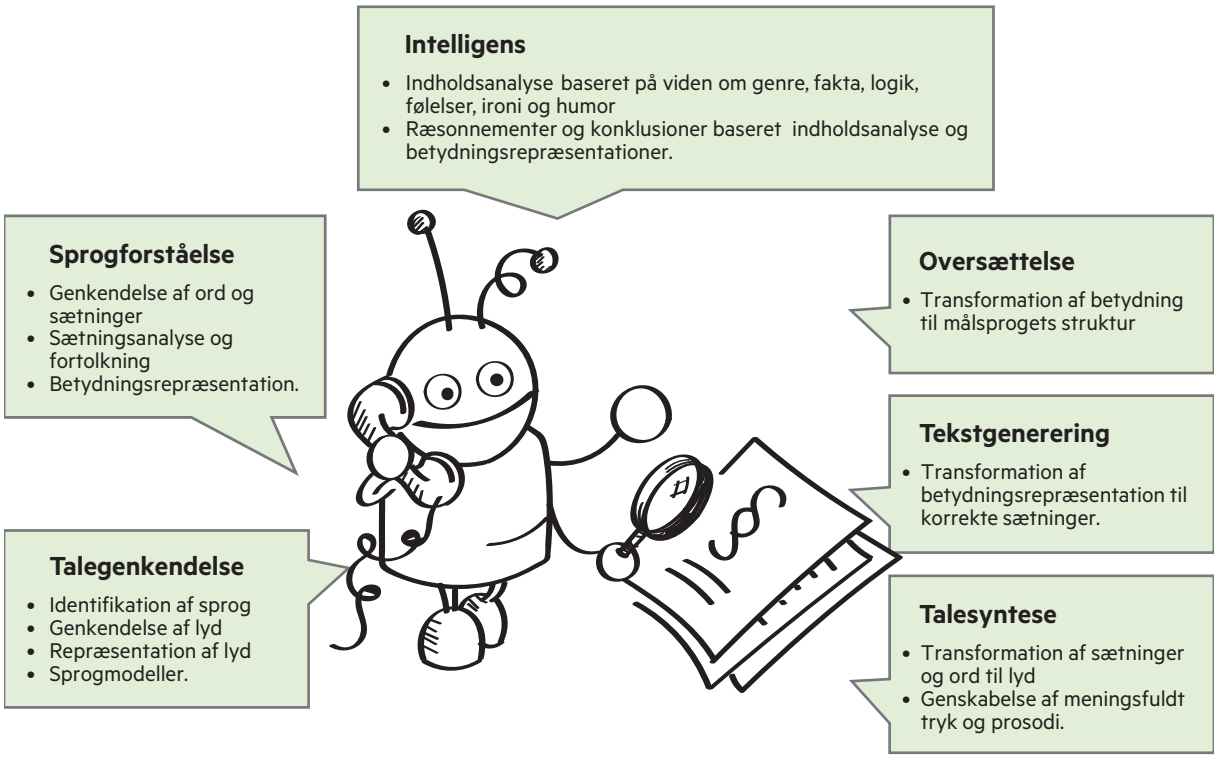
- Danmarks begrænsede størrelse, både som sprogsamfund og som marked
- Det danske sprogs særlige egenskaber (især den komplicerede udtalestruktur)
- Mangel på sproglige basisresurser
- Manglende koordinering af udvikling, distribution og anvendelse af dansk sprogteknologi
- Nedprioritering af forskning og uddannelse inden for dansk sprogteknologi i de seneste år.

På flere områder ligger dansk blandt de lavest rangerende sprog i Europa med hensyn til teknologisk understøttelse, og selv meget mindre sprogsamfund som fx Island og Letland har igennem de seneste år investeret langt mere målrettet i sprogteknologi for deres sprog end Danmark har investeret i dansk sprogteknologi.

2. Hvad er sprogteknologi?

Sprogteknologi er en samlebetegnelse for specialiserede programmer der kan løse sproglige problemstillinger fra simpel stavekontrol til komplekse dialoger mellem mennesker og robotter på et eller flere sprog. Sprogteknologi er specielt gearet til at håndtere de ustrukturerede og ofte flertydige data som menneskeligt talt eller skrevet sprog udgør. Sprogteknologi er en forudsætning for at der kan udvikles kunstig intelligens af høj kvalitet, og bidrager som komponent i alle trin af kommunikationsprocessen. For at man kan bygge et system som en robot som fx kan modtage en besked og udføre en ordre på et givet sprog, skal de grundlæggende data og teknologier være til stede for det pågældende sprog.

Sprogteknologi er en forudsætning for kunstig intelligens



Talegenkendelse: Systemet skal indeholde programmer og sprogmodeller der kan identificere sproget, genkende lydene og omforme dem til en datastruktur som systemet kan arbejde videre med, fx til tekst som brugeren kan se på skærmen eller en til en søgestreng som kan finde informationen på nettet.

Sprogforståelse: Systemet skal indeholde teknologier der kan analysere teksten nærmere og forstå hvad brugeren mener. Er der fx tale om en ordre eller et spørgsmål? Hvad er den mest sandsynlige betydning hvis ordene er flertydige?

Intelligens: Systemet skal kunne foretage de rigtige analyser af vores ytringer, heriblandt identificere særlige markører som humor og ironi, og udlede de fakta der er nødvendige for at agere hensigtsmæssigt. Til dette formål skal systemet kunne inddrage viden om verden, om den betydning vi lægger i begreberne, om den måde vi kategoriserer vores omverden på, om modtageren som dialogpartner og om den aktuelle kontekst i en dialog. Det skal kunne føre meningsfulde samtaler med et minimum af eksplicit information.

Oversættelse: Systemet skal kunne overføre vores kulturelle forståelse til en anden kulturkreds. Til trods for at man efterhånden har opnået særdeles gode oversættelsesresultater med neurale algoritmer, kan de ikke konkurrere i kvalitet og præcision med den menneskelige oversætter, og der er fortsat store udfordringer når algoritmerne er trænet på et emneområde som ligger langt fra den tekst man ønsker at oversætte.

Talesyntese: Hvis systemet skal kunne svare, skal der også være teknologier der ud fra fx en tilbagemelding fra en database kan udforme et forståeligt svar som er passende i den givne situation. Her er det ikke kun gengivelse af ordene i den korrekte rækkefølge der er udfordringen. Det er også placering af pauser og tryk som volder vanskeligheder, da disse meget ofte er meningsbærende i vores kommunikation.

De enkelte teknologier som er nævnt ovenfor, indgår ligeledes som enkeltkomponenter i mange programmer som ikke nødvendigvis involverer kunstig intelligens, som fx dikteringsværktøjer, søgemaskiner, undervisningsprogrammer, tekstanalyse og tekstresumering m.m.

At udvikle sprogteknologi består altså i at kombinere data om selve sproget (såsom statistiske sprogmodeller, ordbøger, termbaser og grammatikker) med modeller af praktiske brugssituationer. I sprogværktøjer til professionelle brugere indgår der ofte en model af et fagområde med en bestemt terminologi og en række betegnelser på produkter, steder, personer eller sagsforhold. Hvis værktøjet er udviklet til at støtte kontakten mellem en sagsbehandler og en borger, en lærer og en elev, en læge og en patient, en politimand og en anholdt m.fl., skal der tilføjes viden om de konventioner, begreber og typiske handlingsforløb som kendetegner situationen.

Udvalget fokuserer først og fremmest på de generelle, basale sproglige komponenter som er hjertet i sprogteknologien, og ud fra hvilke de mere specialiserede applikationer kan blive udviklet. Disse komponenter kan i høj grad genbruges hen over det brede udvalg af sprogteknologier, både dem som allerede er udbredt, og dem som morgendagen vil bringe. En offentlig indsats for at tilvejebringe og vedligeholde disse komponenter er helt central hvis der skal skabes dansk sprogteknologi i verdensklasse.

2.1. Metoder og data

Når man udvikler sprogteknologi, kan man grundlæggende betjene sig af to metoder: regelbaserede eller statistiske.

De **regelbaserede metoder** bygger på nøjagtige beskrivelser af sprogets mindste bestanddele, ordene, og deres kombinationspotentiale, grammatikken. Et regelbaseret system som tager dansk som input, indeholder således en sprogteknologisk ordbog og regler for hvordan systemet skal analysere danske sætningers struktur og fortolke deres indhold.

Statistiske metoder indeholder som udgangspunkt ingen sproglig information, men tager store mængder af data, fx tekst, som input og skaber en statistisk repræsentation af disse data, en såkaldt sprogmodel. Herefter kan den statistiske repræsentation bruges til at analysere nye tekster. Metoden har med stor succes været anvendt inden for automatisk oversættelse, fx Google Translate, EU's eTranslation m.fl., hvor man ved at træne systemer på en stor mængde sætninger og deres oversættelse har kunnet udregne modeller for hvornår et ord i en given kontekst skal oversættes med et andet. Den statistiske repræsentation kan falde sammen med lingvistiske størrelser som ord og ordforbindelser, men gør det ikke nødvendigvis, og det gør det vanskeligt at identificere og reparere fejl.

En særlig form for statistiske systemer er systemer som bruger **neurale metoder,** som er kommet frem i de seneste år, og som fx på oversættelsesområdet har afløst de gamle statistiske oversættere. En af de mest kendte neurale metoder er Word2Vec[1] der klassificerer ord og tekster ved at udregne ligheder mellem ordene ud fra ligheden mellem deres omgivende ord.

Regelbaserede systemer har en styrke i forhold til sprog for hvilke der ikke eksisterer store datasæt, men de er samtidig resursekrævende og somme tider mindre robuste fordi reglerne skal håndkodes af eksperter. Regelbaserede systemer har med succes været anvendt på samisk og grønlandsk, og der findes også kørende regelbaserede systemer for dansk[2].

Statistiske systemer stiller til gengæld store krav til de datasæt som de trænes på, da de i deres output typisk vil genspejle datasættets indhold. Fx kan man risikere at persondata optræder i outputtet hvis det ikke omhyggeligt er anonymiseret. Endvidere er statistiske systemer i højere grad bundet til den teksttype og det

1 https://patents.google.com/patent/US9037464B1/en

2 https://grammarsoft.com/

emneområde som de er trænet på. Det kan således være vanskeligt at nå et godt resultat hvis man anvender et system som er trænet på lovtekster, til at analysere telefonnotater fra kundesamtaler i forsikringsbranchen.

En forudsætning for at man kan udvikle sprogteknologi - uanset om man arbejder statistisk eller regelbaseret - er altså at man har adgang til **sproglige data**. For de regelbaserede systemers vedkommende drejer det sig i høj grad om gode sprogbeskrivelser i form af ordbaser, termbaser og grammatikker, men også om tekstsamlinger med masser af eksempler på det sprog der skal beskrives.

For de statistiske og neurale systemers vedkommende er det fx datasæt bestående af løbende tekst høstet fra internetsider, optagelser af tale, logning af samtaler med en chatbot, en samling af tweets, kommentarer på Facebook og meget mere. Men også de statistiske systemer har brug for eksperternes sprogbeskrivelser i form af annoterede datasæt.

Rå data er datasæt hvor den sproglige information ikke er bearbejdet, som ikke er grundigt strukturerede, og hvor det heller ikke er muligt at filtrere eller sammensætte teksterne efter bestemte kriterier. Metadata kan være tilknyttet de enkelte tekster og giver typisk basale oplysninger. Anvendelse af statistiske eller neurale teknikker på rå data sigter fx imod at sortere tekster efter bestemte kriterier (fx e-mails om bestemte emner) eller at trække information ud (fx information om hvilke virksomheder der oftest nævnes i de danske aviser).

Rå data bruges også i en vis forstand til træning af maskinoversættelse. Dog er der foretaget en vis algoritmisk bearbejdning ved at sætninger i kildetekst og deres oversættelse i målsproget er samlet parvist.

Annoterede datasæt eller **korpusser** består af tekster eller transskriberet tale som er forsynet dels med metadata på tekstniveau, dels med grammatisk opmærkning på ordniveau. Annoteringer på tekstniveau kan fx være domæne (fx "sundhed", "jura" osv.), kilde, dato, overskrift, genrebetegnelse, ophavsperson (evt. med køn, alder og geografisk ophav) og emneord. Grammatiske annoteringer på ordniveau omfatter lemmaform, ordklasse, syntaktisk funktion, ordenes indbyrdes afhængighed, semantisk klasse, anonymiseret ordform, udtale, status som navn, status som del af fast udtryk m.m. Disse annoteringer udgør selve kernen i den værdiberigelse der gør automatisk forståelse af tekster mulig i stor skala.

Det er muligt at inddrage annoteringerne i den statistiske træning af et system og dermed udnytte både den lingvistiske information som man kender fra de regelbaserede systemer, og andre former for annotationer man måtte opmærke data med. Træning på annoterede data kaldes for **dyb læring** (deep learning/supervised learning), mens træning på rå data er **overfladeorienteret læring** (unsupervised learning).

Omfattende korpusser eller datasæt er essentielle **grundresurser** i sprogteknologisk sammenhæng da megen moderne sprogteknologi er baseret på maskinlæringsteknologier, herunder deep learning, der for at fungere godt skal trænes på enorme mængder af tekstmateriale fra et relevant domæne. Korpusser skal vedligeholdes løbende. Ellers vil de hurtigt blive uaktuelle da der konstant kommer mange nye ord og vendinger til.

En stor udfordring for tilvejebringelse af såvel rå som annoterede data er juridiske forhold, nemlig persondataloven og ophavsretsloven. Mange eksisterende korpusser er belagt med ophavsret eller indeholder følsomme data og kan kun stilles til rådighed til forskningsformål. Det er i dag den største forhindring for at der kan opbygges korpusser i den størrelse og domænebredde der skal til for at være til nytte for dansk sprogteknologi. Den mest realistiske måde at løse denne udfordring på er formentlig at indføre undtagelser i ophavsretsloven med henblik på sprogteknologi, at sikre at der indhentes samtykke fra fremtidige datadonorer og at bruge sprogteknologi til effektivt at anonymisere data.

For offentlige institutioners vedkommende giver PSI-direktivet allerede i dag gode muligheder for at stille datasæt frit til rådighed. Der mangler dog den nødvendige viden i institutionerne om vigtigheden af at dette sker, og at ansvaret for at indsamle data og informere institutionerne placeres i et centralt organ. Dette behov understreges ligeledes i Rigsrevisionens seneste beretning Åbne data[3].

3 http://www.rigsrevisionen.dk/publikationer/2019/122018/

2.2. Taleteknologi

Talesproget er menneskers mest naturlige kommunikationsmiddel. Vi er født til at tale, og stort set alle mennesker udvikler et effektivt og udtryksfuldt talesprog længe inden de kommer i skole og lærer at skrive. Mange ellers normalt udviklede sprogbrugere har svært ved at lære at anvende skriftsproget. Nyere undersøgelser konkluderer at omkring hver ottende dansker er 'funktionel analfabet' (ref.)[4], dvs. nok i stand til at læse og skrive, men kun med så stort besvær at de i praksis næsten aldrig anvender skriftsproget. Dette er en stor udfordring i næsten enhver professionel sammenhæng, ikke mindst i de tertiære erhverv, og det medfører et betydeligt værditab for både den enkelte og for samfundet. Men også mennesker uden kommunikationsvanskeligheder kan opleve tastaturet som en barriere ved betjening af software man ikke er fortrolig med, såsom offentlige informationstjenester, elektroniske formularer, portaler, søgemaskiner, skoletjenester, for slet ikke at tale om hjemmets og arbejdspladsens hardware. Her kan de fleste let beskrive deres ønsker mundtligt, men kun med besvær omsætte dem til tastetryk.

Derfor er talesproget som it-modalitet blevet et globalt satsningsområde gennem de seneste ti år. Ved hjælp af taleteknologi kan man direkte betjene software og maskiner med stemmen.

Talesyntese

Talesyntese (TTS, Text-to-Speech) sætter en computer i stand til at efterligne et talende menneske. De fleste kender kunstige danske talestemmer fra fx GPS'er og offentlige transportmidler. I kombination med talegenkendelse udgør talesyntese stammen i taleassistenter (fx Siri og Google Assistent), og man forventer at denne type dialogsystemer vil spille en stadigt større rolle i Danmark, både privat og professionelt i de kommende år. Allerede fra 2019 ventes de største bilmærker at tilbyde dansktalende dialogsystemer som ekstraudstyr i en lang række bilmodeller.

Især syns- og talehandikappede har haft fordel af talesyntese, som har bragt bøger, artikler og tidsskrifter inden for den blindes hørevidde og givet den multihandikappede en talestemme. Den engelske fysiker Steven Hawking var en verdenskendt fortaler for talesyntesen og brugte den som sin omverdenskontakt helt til sin dødsdag i marts 2018.

Talesyntesen er generelt blevet mødt med skepsis herhjemme. De tidlige danske syntesestemmer lød kunstigt, især når de skulle foregive at tale 'almindeligt dansk'. De seneste par år er kvaliteten dog forbedret markant, ikke mindst med skiftet til algoritmer baseret på neurale net. Nu er synteserne, for dansk og mange andre sprog, blevet så naturtro at man i kortere forløb ikke altid kan afgøre om stemmen tilhører et menneske eller en computer. Dette gælder dog ikke i dialoger mellem mennesker og computere, hvor maskinens kunstige udtale og generelt ringe sprogbehandling stadig er en betydelig barriere.

Talegenkendelse

I disse år er talegenkenderen den mest omtalte sprogteknologi. Talegenkendelse (ASR, Automatic Speech Recognition) er i stand til at registrere menneskelig tale og genkende de sagte ord. I kombination med intelligent software kan en talegenkender fungere som en sekretær som tager imod et diktat og direkte omsætter det til en færdigformateret tekst. Denne teknologi er vidt udbredt i Danmarks offentlige sektorer, både i regioner (fx diktering af patientjournaler), kommuner (fx sagsbehandleres rapportskrivning) og Folketinget (afskrifter af debatten til Folketingstidende).

Hospitalssektoren har brugt talegenkendelse i årevis, særligt inden for områder som patologi og kirurgi, hvor det er af stor værdi for personalet at kunne betjene computerne uden at røre et tastatur. Læger og assistenter kan i stedet diktere direkte ind i sygejournalen. Derved mindskes både tidsforbruget og risikoen for fejl, som ellers er en alvorlig faktor når diktater aflyttes og nedskrives af lægesekretærer.

En række nye erhverv og organisationer tester i disse år talegenkendelsens muligheder, herunder politi og militær, byggeriet, produktionserhvervene, biblioteker, privatpraktiserende advokater og tandlæger og mange andre. Mange rapporterer at mulighederne er oplagte, men at kvaliteten af de tilgængelige danske talegenkendere er for lav til at Danmark fuldt ud kan udnytte teknologiens potentiale. At udvikle dansk talegen-

4 http://denstoredanske.dk/Erhverv,_karriere_og_ledelse/P%C3%A6dagogik_og_uddannelse/P%C3%A6dagogik,_didaktik_og_metodik/analfabetisme

kendelse i høj kvalitet kræver først og fremmest tilgang til danske taledata af høj lingvistisk kvalitet, resurser som er kostbare at etablere, men som til gengæld i høj grad kan genanvendes.

2.3. Tekstanalyse

For de fleste voksne mennesker er læsning en dagligdags aktivitet som ikke volder besvær. Læsningens delprocesser er blevet automatiseret, og man er ikke længere bevidst om styringen af sine øjenbevægelser, registreringen af ordmellemrum, genkendelsen af tegn, oversættelse af tegngrupper til genkendelige ord, opsamlingen af ord til fraser, afkodning af frasernes grammatik og betydning. Ordene synes tværtimod at flyde ind i bevidstheden lige så enkelt som luften flyder ned i lungerne.

Computeren har ikke samme flair for sprog. Hver eneste af de nævnte delprocesser er en udfordring for den automatiske tekstanalyse og kræver et højt specialiseret sprogteknologisk værktøj. At indlæse en tekst kræver **datavask**, dvs. fjernelse af fx personoplysninger og diverse formateringskoder i teksten (deformatting), at identificere tegngrupperne kræver **ordidentifikation** (tokenization), at omsætte dem til genkendelige ord der kan slås op i ordbasen kræver **lemmatisering** (lemmatization), at forberede dem til sætningsanalyse kræver **opmærkning** (tagging) og at samle dem til en forståelig sætning kræver **sætningsanalyse** (parsing). Først derefter kan sætningen bruges som input til et søgesystem, omsættes til tale eller oversættes til et andet sprog.

I sprogteknologiens barndom var kun de lavere tekstfunktioner inden for rækkevidde. Tekstbehandlerens stavekontrol byggede på simpel ordgenkendelse og kunne identificere fejlstavede ord som "væm" og "våres", men ikke kontekstafhængige fejl som "et dejligt vær" og "jeg køre hjem". Med udviklingen af bedre taggere og parsere har vi fået langt bedre korrekturprogrammer som endda kan anvise alternative formuleringer og kommentere på stil og indhold i længere tekstafsnit.

De seneste år har statistisk baserede sprogmodeller i nogen grad udkonkurreret de traditionelle regelbaserede metoder. Det gælder fx Word2vec-modellen som ligger til grund for algoritmer som automatisk finder mønstre i store mængder tekst og rekonstruerer ordenes lingvistiske kontekst (word embeddings)[5]. Ord der optræder i samme kontekst, viser sig typisk at være beslægtet med hinanden og kan derfor klassificeres ud fra konteksten. Kvaliteten af det output algoritmen leverer, afhænger imidlertid meget af det korpus som lægges til grund, og er også meget følsomt i forhold til sjældne eller ukendte ord, og der arbejdes derfor flere steder på at opnå bedre og mere pålidelige resultater ved at kombinere Word2vec med sprogteknologiske ordbøger, terminologidatabaser, ordnet og lign.

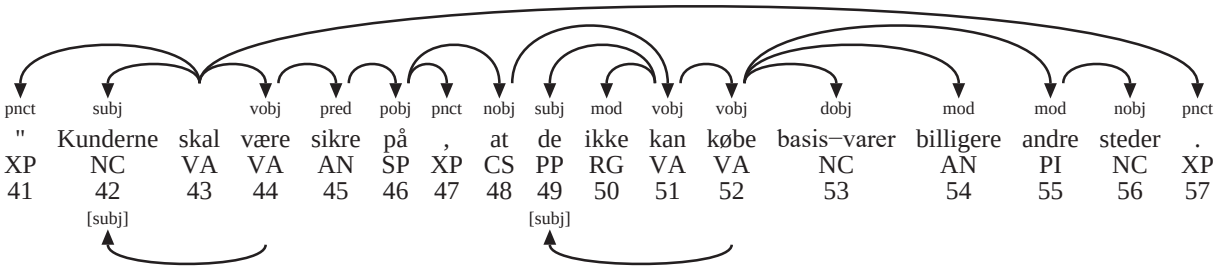
Med kraftige computere trænet på store tekstmængder kan man opnå en kompetent tekstanalyse som ofte kan konkurrere med menneskers. For eksempel anvendes sentimentanalyse til at vurdere teksters emotionelle indhold og synspunkter, retorisk strukturanalyse til at kortlægge argumentationen i sagsakter, samt emneklassifikation til at resumere og kategorisere videnskabelige artikler. Tekstanalyse anvendes typisk af analysebureauer, medier, marketing, politiske organisationer og mange andre virksomheder som er specialiseret i vidensindsamling.

Opmærkning

Opmærkning af tekster med lingvistiske informationer af høj kvalitet er nødvendig i maskinlæringsteknologier for at opnå sprogmodeller med den optimale kvalitet. Der er derfor behov for en række kvalitetskorpusser med kontrolleret opmærkning – såkaldte **guldstandarder** – som kan danne et solidt grundlag for træning af sprogteknologiske komponenter i danske sprogteknologiske virksomheder.

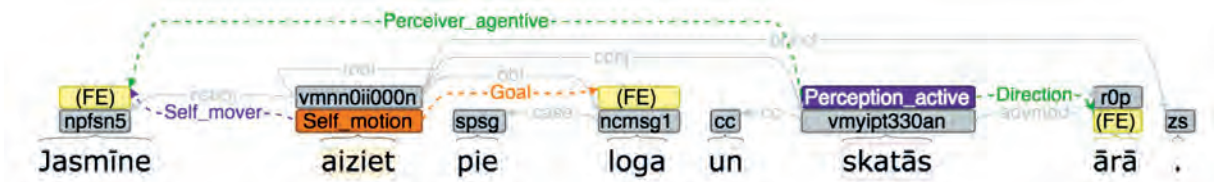
Det gælder både syntaktisk information om ordklasser (part of speech), og syntaktiske relationer som subjekt, objekt mv., og semantisk information om ords betydning, deres betydningsmæssige relationer, referencer til andre ord og deres valør, fx om de er positivt eller negativt ladede. Det gælder også opmærkning af talesprog især med henblik på talegenkendelse.

5 Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". https://arxiv.org/abs/1301.3781



Eksempel på syntaktisk opmærkning fra DDT (Danish DependencyTreebank).

Kvalitetsopmærkning af tekst af denne type er typisk håndarbejde. Der findes derfor ikke mange guldstandarder der har opmærkning af høj kvalitet, og de er typisk ret små, fx dækker Danish Dependency Treebank som er udviklet ved CBS, ca. 100.000 ord i løbende tekst[6], og SemDaX-korpusset som er udviklet ved Center for Sprogteknologi ved Københavns universitet dækker, 90.000 ord[7].



Eksempel på opmærket tekst fra det lettiske framenet[8].

Guldstandarder er notorisk dyre og tidkrævende fordi kvaliteten skal være i absolut top. For at sikre en konsistent opmærkning bruges fx ofte flere personer til at opmærke den samme tekst hvorefter man sammenligner deres resultater og udligner forskelle, hvilket driver prisen og også tidsforbruget i vejret.

Selv om antallet af ord som er opmærket til guldstandard, er lille, kan opmærkning af mere materiale ofte gøres hurtigere og billigere med brug af forskellige maskinlæringsteknikker og interaktive metoder (hvor mennesker gennemgår og retter maskinens opmærkningsforslag), når først guldstandarden foreligger.

2.4. Sprogforståelse

Der er stadig stor diskussion blandt psykologer og kognitionsforskere om hvilken form for 'forståelse' en maskine kan opnå. Den klassiske Turing-test er kendt som en metode til at afgøre, om en maskine udviser intelligens på menneskeligt niveau. Turing-testen er bestået hvis det menneske som interagerer med maskinen, ikke er i stand til at afgøre om hun faktisk kommunikerer med en maskine eller et andet menneske. Men testen siger ikke noget om hvorvidt systemet er intelligent, men snarere noget om hvad det interagerende menneske opfatter som intelligens.

En vis form for forståelse ses også i beslutningsstøttesystemer hvor systemets evne til at ræsonnere og drage logiske slutninger kan hjælpe et menneske med bedre at overskue komplekse sammenhænge ud fra givne data, fx en læge med at stille en diagnose på basis af en række symptomer.

Den praktisk orienterede sprogteknologi har identificeret en lang række områder hvor selv en overfladisk efterligning af menneskers sprogforståelse giver en klart forbedret brugeroplevelse. I USA har mange banker fx løst problemer med ventetid ved at erstatte menneskelige telefonoperatører med automatiske agenter, og kundeundersøgelserne viser stor tilfredshed. Til det engelske sprog findes der allerede et større udbud af automatiske sekretærbots der tager sig af enklere kontorfunktioner som lokalebestilling, opdatering af mødekalender og fremfinding af dokumenter. Markedet for sprogteknologi til intelligent interaktion med menne-

6 http://www.buch-kromann.dk/matthias/ddt1.0/

7 https://cst.ku.dk/english/projekter/projekter-afsluttet/semantikprojekt/corpus/

8 https://www.clarin.eu/blog/clarin-latvia-presents-latvian-framenet

skelige brugere er inde i en rivende udvikling. Der er en betydelig kreativitet i markedet, og ingen kan forudse hvem og hvad vi kan kommunikere med om ti år.

Den største udfordring omkring sprogforståelse er imidlertid systemernes manglende viden om betydningen af de data de arbejder med, manglende viden om verden og manglende viden om de kulturelle konventioner der opstår i et samfund. Derfor er adgang til strukturerede og verificerede data om fx dansk kulturarv, danske geografiske og samfundsmæssige forhold, almensproglige ords og termers betydning og konventioner for deres brug en vigtig forudsætning for at man i fremtiden kan bevæge sig videre i retning mod en dybere sprogforståelse.

Fx forsøger adskillige systemer at eksperimentere med at inddrage encyklopædisk viden, fx data fra Wikipedia. Her støder man imidlertid ofte på det problem at det er ret tilfældigt hvilke oplysninger der bliver lagt ind for de enkelte kultur- og samfundsområder, og at kvaliteten af de indtastede oplysninger er meget svingende, hvilket medfører at sådanne svagheder så igen afspejler sig i det resulterende system.

Adgang til kuraterede encyklopædiske data af høj kvalitet vil have betydning for at man med tiden kan opnå mere avancerede og pålidelige sprogforståelsessystemer, men med det stade som systemerne har nu, vil der stadig være lang vej før denne type sprogforståelse vil være mulig. Store nationale kulturarvsprojekter som fx Gyldendals Encyklopædi og Trap Danmark er særdeles interessante og relevante resurser af høj kvalitet som bør være tilgængelige som korpusser med henblik på udvikling af sprogteknologi. Institutionerne bag disse produkter har tilkendegivet at de ikke vil være afvisende over for at stille resurserne til rådighed til sprogteknologiske formål.

Helt afgørende for at komme videre hen imod en dybere sprogforståelse er det imidlertid at der foreligger beskrivelser der sætter systemerne i stand til at udnytte viden om de enkelte ords og termers betydning og indbyrdes relationer.

Der findes allerede en række vigtige resurser der systematisk og detaljeret beskriver danske ords betydning og kombinationspotentiale, først og fremmest baseret på mangeårige samarbejdsprojekter mellem Det Danske Sprog- og Litteraturselskab og Center for Sprogteknologi ved Københavns Universitet. Arbejdet bygger især på betydningsoplysningerne i Den Danske Ordbog og er i overvejende grad finansieret af forskningsråds- og fondsmidler. Det gælder resurserne FrameNet, der beskriver semantiske scenarier for 5300 verber og 6490 verbalsubstantiver, SemDax, hvor ordenes betydninger og relationer er opmærket i løbende tekst, og DanNet, der beskriver ordenes relationer til andre ord[9]. Den 1.2.2019 er det med en bevilling på 2 mio. kr. fra Carlsbergfondet blevet muligt at udvide dækningsgraden for DanNet som pt. omfatter 66.308 begreber.

Med midler fra Kulturministeriet, som har givet bevillinger til Den Danske Ordbog, fra forskningsrådene, som har givet midler til DanNet og FrameNet igennem årene, og fra Carlsbergfondet, som ligeledes i flere omgange har understøttet udviklingen af flere af resurserne, er der således allerede etableret et godt fundament for det videre arbejde med sprogforståelse.

De to institutioner har planer om at koble deres forskellige resurser sammen til en samlet ordbogsresurse. En sådan samlet resurse vil – især hvis den også kobles sammen med en udtaleresurse og en termbank – kunne udgøre en vigtig basisbyggesten i en fri dansk sprogresurse der kan danne grundlag for systemernes sprogforståelse. Dele af DanNet og FrameNet er endvidere koblet til andre sprog, hvilket også vil gøre en sådan resurse anvendelig i flersprogede applikationer. Begge institutioner samarbejder endvidere med tilsvarende institutioner på europæisk plan i EU-projektet eLEXIS på dels at udvikle fælles værktøjer og standarder for leksikalske resurser og dels at sammenkoble ordoplysninger på tværs af sprogene i hele Europa[10].

Med udgangspunkt i de allerede eksisterende sproglige beskrivelser vil det blive muligt at kvalitetsopmærke tekst med den semantiske information der findes i de leksikalske resurser - en guldstandard inden for dansk semantik - til brug for maskinlæring.

9 http://www.lrec-conf.org/proceedings/lrec2018/pdf/586.pdf

10 https://elex.is/

2.5. Terminologi, vidensmodellering og it-arkitektur

En term er et fagudtryk. Alle som beskæftiger sig med et bestemt fagområde, hvad enten man er studerende, underviser eller medarbejder i en privat eller offentlig virksomhed, har brug for at kende de fagudtryk der bruges når man taler om sit fag. En stor del af undervisningen i uddannelsessystemet bruges på at introducere og forklare de udtryk man bruger når man taler om et genstandsfelt eller fagområde, fx klima eller kemi.

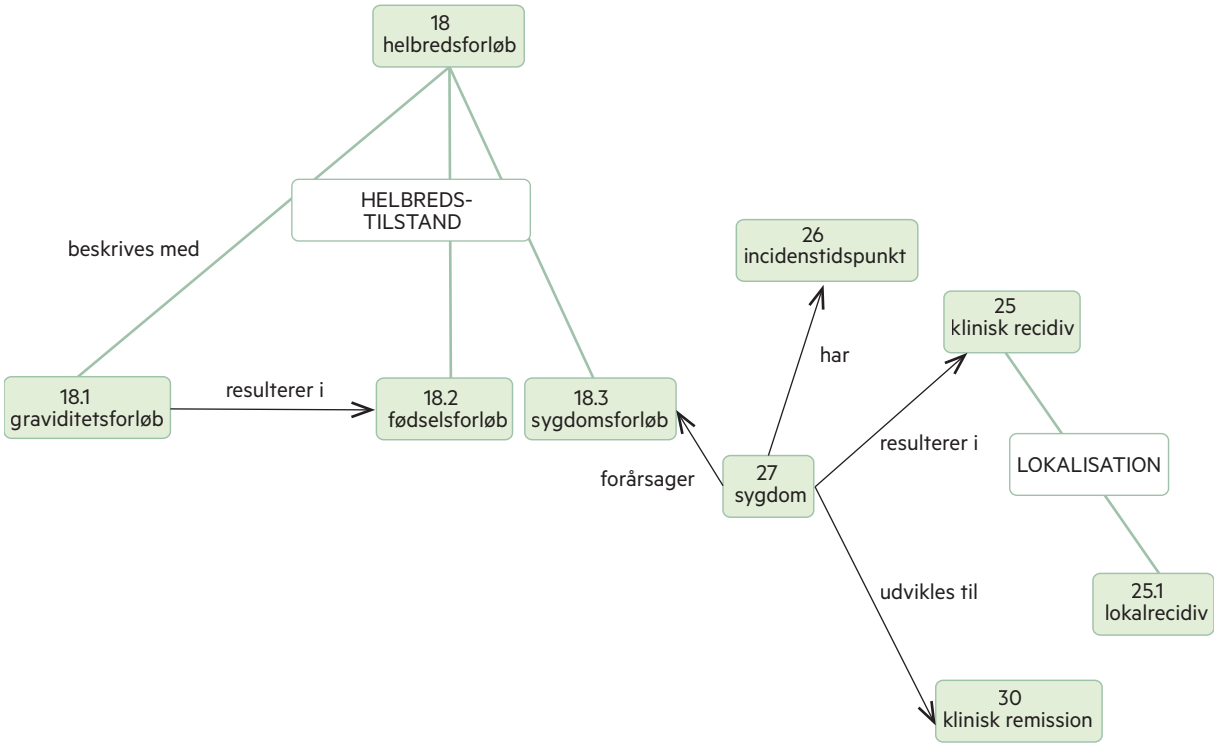
Et fagområdes terminologi dækker de fagudtryk man bruger i det pågældende fag. Fagudtrykkene kan være beskrevet på forskellige måder og med forskellig detaljeringsgrad. En typisk måde at beskrive fagudtryk på er fagordbøger, fx juridiske, tekniske eller økonomiske ordbøger. Her er fagudtrykkene alfabetisk ordnet. En anden måde at beskrive fagudtryk på er at ordne dem tematisk, det kender man fx fra bibliotekernes og museernes emneklassifikationer, hvor de enkelte emner er opdelt i underemner.

Vidensmodellering

Arbejdet med terminologi tager udgangspunkt i det sproglige udtryk, dets begrebsindhold og begrebets relationer til andre begreber. En begrebsorienteret terminologi tegner typisk et billede af verden således som eksperter opfatter den. Når relationerne mellem begreber opstilles i diagrammer eller modeller, taler man om vidensmodellering.

Et eksempel er Sundhedsvæsenets Begrebsdatabase (NBS)[11] som har til formål at skabe en fælles forståelse for sundhedsfaglige begreber på tværs af sundhedsvæsenet. Heri beskriver Sundhedsstyrelsen centrale begreber og deres indbyrdes relationer således at det bliver tydeligt for medarbejdere i sundhedsvæsenet og deres samarbejdspartnere, fx leverandører af it-systemer, hvordan begreberne skal forstås. Her fokuseres der på relationerne mellem begreberne, fx hvad der er over- eller underordnet, men også på andre typer relationer, fx årsagssammenhænge, tidsforløb, egenskaber og lign.

I nedenstående eksempel ses modelleringen i NBS af begrebet helbredsforløb hvor det bliver tydeligt at graviditetsforløb og fødselsforløb er helbredstilstande som ikke betragtes som sygdomsforløb. Et sygdomsforløb forårsages nemlig af sygdom, og sygdom er så nærmere beskrevet som noget der har et bestemt inci-



Uddrag af begrebssystem fra Sundhedsvæsenets Begrebsdatabase (NBS).

11 https://sundhedsdatastyrelsen.dk/da/rammer-og-retningslinjer/om-terminologi/nbs

denstidspunkt (det tidspunkt hvor en diagnose stilles første gang), og som noget der kan forsvinde (klinisk remission) eller komme igen (klinisk recidiv) [12].

Bag hvert enkelt begreb i begrebsdatabasen ligger en definition af begrebet og en angivelse af evt. synonymer samt links til relaterede begreber, her fx begrebet *klinisk recidiv*.

Opslag i Sundhedsvæsenets Begrebsdatabase (NBS).

En ny faglig eller administrativ medarbejder, en borger eller en leverandør af et it-system vil således til enhver tid ved at slå op i NBS kunne finde frem til hvad der menes med termen når den optræder i en given sammenhæng, fx en arbejdsbeskrivelse, en patientjournal eller en kravspecifikation. Tilsvarende findes der begrebsdatabaser hos Socialstyrelsen[13].

Vidensmodellering, terminologi og it-arkitektur

Forståelse af begreberne og deres relationer har stor betydning når vi ønsker at bruge it-systemer til at understøtte arbejdsprocesser. Det kan være elektroniske patientjournaler, det kan være databaser i Danmarks Statistik eller kommunernes dokumentstyringssystemer.

Mange it-projekter har været sat i værk i tidens løb med skiftende succes, og i dag er en af de store udfordringer at integrere de forskellige systemer med hinanden. De mange tusinde offentlige it-systemer taler ikke et fælles sprog, og der har ikke tidligere været en fællesoffentlig plan for, hvordan it-systemerne sikkert og effektivt skal kunne udveksle data og indgå i sammenhængende processer. Datadeling er blevet aftalt fra gang til gang, og resultatet er "spaghetti-integrationer", der har været dyre at udvikle og nu er svære og endnu dyrere at vedligeholde.[14]

Derfor har regeringen, kommunerne og regionerne aftalt at der som en del af den fællesoffentlige digitaliseringsstrategi 2016-20 skal etableres en fællesoffentlig arkitektur for sikker og effektiv deling af data og tværgående processer. Det begrundes bl.a. således:

> *Borgerne og virksomhederne skal opleve, at behandling og service på tværs af flere myndigheder hænger bedre sammen. De samme data skal ikke indsamles flere gange, da dette koster penge og giver alle mere besvær. Myndighederne skal kunne trække på hinandens viden og kunne samarbejde til gavn for borgere og virksomheder.*

12 https://sundhedsdatastyrelsen.dk/da/rammer-og-retningslinjer/om-terminologi/nbs/om-arbejdsgrupperne/klinisk_administrative-begreber

13 http://www.socialebegreber.dk/ NB! Den har også gode publikationer med illustrationer af forholdet mellem begreber og it-arkitektur.

14 https://arkitektur.digst.dk/mandat-og-styring/hvidbog-om-faellesoffentlig-digital-arkitektur s. 3

Som led i aftalen har Digitaliseringsstyrelsen opstillet tydelige og bindende krav til hvordan offentlige institutioner skal gribe udvikling af nye it-projekter an[15].

Der er opstillet en række overordnede principper for arkitekturarbejdet:

1. Arkitektur styres på rette niveau efter fælles rammer
2. Arkitektur fremmer sammenhæng, innovation og effektivitet
3. Arkitektur og regulering understøtter hinanden
4. Sikkerhed, privatliv og tillid sikres
5. Processer optimeres på tværs
6. Gode data deles og genbruges
7. It-løsninger samarbejder effektivt
8. Data og services leveres driftssikkert.

Især princip 3 og 6 peger på nødvendigheden af at skabe et fælles sprog og en entydig forståelse af begreber:

Princip 3:
Arkitektur og regulering understøtter hinanden. Projekters arkitekturleverancer bidrager til at sikre, at lovgivning og anden regulering overholdes. Modsat skal arkitekturen understøtte at ny lovgivning og anden regulering er digitaliseringsklar. Dette kan ske ved at identificere komplicerede regler i lovgivningen eller ved at **fastlægge fælles begreber og bidrage til en entydig forståelse af disse på tværs af lovgivningen.**

Princip 6:
Gode data deles og genbruges. Data er en ressource som gennem deling og genbrug anvendes til at skabe værdi for borgerne og virksomhederne og til at skabe sammenhæng i den offentlige sektor. **Begreber og data beskrives ensartet, så de kan genbruges og der sikres tilstrækkelig kvalitet i data til de væsentlige anvendelser af data.**

For hvert arkitekturprincip er der opstillet en række arkitekturregler, der nærmere specificerer hvordan man bedst lever op til principperne.

Arkitekturregel 6.2:
Anvend fælles regler for dokumentation af data For at fremme genbrug af data beskrives data og begreber efter fælles regler. Det er nødvendigt for at sikre, at data forstås korrekt og passer sammen, når de anvendes på tværs af myndighedernes forskellige processer og it-systemer. Det betyder at:

- Projekter anvender de fællesoffentlige regler for begrebs- og datamodellering til at beskrive **den semantiske betydning og modellering af data.** Reglerne understøtter, at der skabes sammenhæng fra begreberne i lovgivningen til data, der udstilles via et it-systems snitflader.
- **Projekter beskriver deres data og begreber så fyldestgørende, at de kan forstås og genbruges i andre sammenhænge.**

15 https://arkitektur.digst.dk/mandat-og-styring/hvidbog-om-faellesoffentlig-digital-arkitektur s. 4 ff.

Arkitekturregel 6.4:
Udstil oplysninger om datakilder, begreber og datamodeller Beskrivelser af datakilder, begreber og datamodeller udstilles således, at myndigheder og private kan få indsigt i, hvilke data offentlige myndigheder har og dermed vurdere potentielle muligheder for genbrug. Det betyder at:

- **Beskrivelser af datakilder, begreber og datamodeller udstilles efter fælles standarder, fx på myndighedens hjemmeside eller i et fælles katalog.**

Arkitekturreglerne anbefaler at der som det allerførste trin i processen udarbejdes definitioner af begreber eller begrebsmodeller, og at det angives om definitionerne er taget fra en fælles model, eller om de er udviklet og defineret internt i organisationen. Udgangspunktet er således modeller af fagpersonernes forståelse af hvordan processerne er eller bør være baseret på de centrale begreber der anvendes i praksis, mens de egentlige datamodeller først udvikles i næste trin.

Deling af begrebs- og datamodeller

En helt central brik i den fællesoffentlige it-arkitektur er deling af de begrebs- og datamodeller som skabes hos de forskellige aktører, så de kan genbruges af andre til forskellige formål. Den overordnede vision er at begrebs- og datamodeller kun skal være defineret og udstillet ét sted. Dertil er vi langt fra nået endnu. Der foregår et stort arbejde med datamodellering rundt omkring i styrelserne og kommunerne, men det foregår uafhængigt af hinanden og uden en overordnet styring. Det er helt bevidst fordi man satser på at der udarbejdes delmodeller efterhånden som behovene for dem opstår, og at delmodellerne kan genbruges og sættes sammen med andre modeller efterhånden som de godkendes og udstilles i en fælles portal.

Begrebs- og datamodeller deles i Digitaliseringsstyrelsens modelkatalog[16], som bl.a. indeholder internationale standarder og grunddatamodeller som modellerer de frie grunddata der distribueres via Datafordeleren som er en webtjeneste der drives af Styrelsen for Dataforsyning og Effektivisering. Modellerne indeholder i mange tilfælde gode og klare definitioner af de begreber og sammenhænge de modellerer.

Der findes pt. kun ganske få fora hvor de sproglige udfordringer med begrebs- og datamodellerne diskuteres. Et af dem er netværket FORVIR (Forum for vidensmodellering i offentligt regi) som består af repræsentanter fra forskellige offentlige institutioner, dvs. terminologer og vidensarkitekter der jævnligt mødes for at dele erfaringer og ideer i forhold til den fællesoffentlige it-arkitektur.

Begrebsarbejde, kommunikation og den politiske proces

Digitaliseringsstyrelsens arkitekturregler giver god mening. Det er i sidste ende menneskers forståelse af hvordan tingene hænger sammen, der skal afspejle sig i systemarkitekturen, og det er mennesker der skal indtaste data i systemerne, og mennesker der skal fortolke de data der trækkes ud af systemerne.

Den fællesoffentlige digitaliseringsstrategi kræver endvidere at der kommer sammenhæng mellem lovgivningen og arkitekturleverancerne, dvs. den skal bidrage til at sikre at lovgivning og anden regulering overholdes, og på den anden side understøtte at ny lovgivning og anden regulering bliver digitaliseringsklar. Det betyder også at de anvendte begreber skal formidles til de politikere og jurister der laver lovene, og i sidste ende til borgerne som skal forholde sig til lovene og bliver berørt af dem. Via klassificeringssystemer som fx FORM (Den Fællesoffentlige Referencemodel) er der allerede taget et stort skridt i retning af at beskrive de offentlige opgaver og deres lovgrundlag på en ensartet, overskuelig og gennemskuelig måde, og det er bestemt et godt skridt på vejen.

Men andre begreber som også har betydning for vores samfund, kan i dag ikke slås op et fælles sted. Ganske vist findes ord som *arbejdsprøvning*, *indlæggelse* og *efterkommer* beskrevet i ordbøgerne, men det er typisk den meget overordnede, almene forståelse af begreberne der bliver afspejlet der. Der er ingen fælles indgang til hvordan de begreber bruges rundt omkring i love og bekendtgørelser, i Socialstyrelsens begrebsdatabase, i Sundhedsstyrelsens begrebsdatabase eller i redegørelser fra Danmarks Statistik og mange andre steder, og der

16 https://arkitektur.digst.dk/node/610

er ingen overordnet styring af hvilke fagområder der skal modelleres, eller kontrol af om definitionerne stemmer overens på tværs af områderne. Behovet for arkitekturregler understreger tydeligt at offentlige institutioner først lige er begyndt at forstå sammenhængen imellem den menneskelige forståelse af begreberne og den måde de defineres på i de it-systemer der styrer vores hverdag. De institutioner der har satset på begrebsarbejdet igennem længere tid, kan efterhånden mærke de økonomiske og arbejdsmæssige fordele det medfører.

En fælles termbank der giver alle adgang til de definitioner af ord og begreber som udarbejdes rundt omkring i offentlige institutioner i forbindelse med den offentlige digitaliseringsstrategi, kunne nyttiggøres på mange måder i den daglige kommunikation og ikke blot lette forståelsen blandt institutionernes medarbejdere, men også mellem institutionerne, politikere, journalister og borgere.

Arbejdet med beskrivelsen af begreberne er allerede i fuld gang. Det der mangler, er et sted hvor de samles og gøres tilgængelige for alle, fx en dansk termbank.

Termbankens betydning for sprogteknologi og kunstig intelligens

Adgang til begrebsmodeller for bestemte fagområder kan få stor betydning for udvikling af dansk sprogteknologi og understøtte udvikling af mange former for applikationer inden for de forskellige fagområder.

- På ordniveau: fx adgang til det centrale faglige ordforråd og inkl. systematisk betydningsangivelse
- På sætningsniveau: fx adgang til relationer mellem begreber til udvikling af sprogforståelse
- På tekstniveau: fx automatisk tekstklassifikation på basis af domænemodeller
- På domæneniveau: fx understøttelse af automatiske ræsonnementer og intelligent dialog.

Der er nogle grundlæggende forskelle mellem en dansk ordbogsresurse og en dansk termbase som gør at de to resurser som udgangspunkt bør holdes adskilt. For det første er indholdet meget forskelligt: Ordbøgerne indeholder det almensproglige ordforråd, og mange af ordene har flere betydninger, mens termbaserne indeholder fagordforrådet og her er ordene typisk entydige – i hvert fald inden for de enkelte fagområder. For det andet er produktionsprocessen ligeledes forskellig: Ordbøgerne bliver udviklet og opdateret af leksikografer og lingvister, mens termbaserne bliver opdateret af fageksperter som ikke nødvendigvis har en sproglig baggrund, og (som vist ovenfor) ofte også i tilknytning til udvikling af offentlige it-projekter. For det tredje er beskrivelsesmåderne forskellige. Ordbøgerne tager typisk udgangspunkt i det enkelte ord, mens termerne typisk er knyttet til begreber, dvs. der kan være flere forskellige termer der betegner det samme begreb.

Det er imidlertid hensigtsmæssigt at ord og termer gøres tilgængelige samlet således at man kan finde de leksikalske enheder uanset i hvilken database de er beskrevet, og at der oprettes links mellem beskrivelserne, således at man fx kan se hvordan et givet ord er beskrevet som term og som alment udtryk, idet der typisk vil være en del overlap imellem de to beskrivelsesformer.

2.6. Automatisk oversættelse

Afhængigt af teksttypen foretages oversættelsen automatisk eller semiautomatisk via en oversættelseshukommelse i interaktion med en oversætter eller translatør. Brugen af oversættelseshukommelse er langt den mest udbredte metode, men automatisk oversættelse vinder mere og mere frem.

Oversættelseshukommelser

En oversættelseshukommelse er en database hvor sætninger og deres oversættelse er lagret parvis. Når en ny tekst skal oversættes, kan tidligere oversatte sætninger fremfindes via statistisk match med den aktuelle sætning og tilbydes som oversættelsesforslag som derefter tilrettes manuelt. Den oversatte kontrollerede sætning gemmes herefter i databasen. I mange tilfælde kan systemet kobles med et automatisk oversættelsessystem som kommer med forslag til oversættelse når der ikke findes et match i databasen. Hvis oversættelseshukommelser kombineres med en flersproglig terminologidatabase, kan det yderligere øge kvaliteten i oversættelsen.

Oversættelseshukommelser er særdeles interessante som sproglige resurser i mange sammenhænge fordi de indeholder store mængder parrede sætninger på både kildesprog og målsprog, ofte struktureret efter kunder eller faglige domæner. Da oversættelserne ofte er foretaget af professionelle sprogfolk, er hukommelserne

typisk af høj kvalitet og dermed særdeles velegnet til maskinlæringsopgaver, fx med henblik på opbygning af sprogmodeller eller oversættelsessystemer. EU's statistisk/neurale maskinoversættelsessystem eTranslation er fx trænet på oversættelseshukommelser som er produceret af EU's sprogtjenester.

Automatisk oversættelse

Automatiske oversættelsestjenester på nettet som Google Translate eller Microsoft Bings Translator har eksisteret i mange år og er ofte blevet kritiseret for deres kvalitet især når det gælder oversættelse til og fra dansk. Den teknologiske udvikling, først brugen af statistisk oversættelse og siden brugen af neurale algoritmer, har i de seneste 2 år øget kvaliteten betydeligt, og begge systemer anvendes i dag i praksis i stor stil og tilbydes både online og som api.

Systemerne trænes på tekster som virksomhederne høster fra nettet, og via crowd-sourcing hvor brugere rapporterer fejl og indsender korrigerede tekster for at forbedre systemet. Denne fremgangsmåde fungerer fint inden for det almene ordforråd og på specialiserede områder hvor der er meget tekst tilgængelig, men mindre godt inden for specialiserede områder eller i situationer hvor tekster kun findes i en styrelses, en kommunes eller en virksomheds interne dokumentarkiver.

Pt. er oversættelsens kvalitet bedst mellem sprogpar som ofte forekommer sammen på nettet. Det er tydeligt at kvaliteten falder allerede når man bevæger sig fra dansk-engelsk til dansk-tysk, og kvaliteten falder yderligere når man anvender andre sprog, fx hvis man har behov for at få oversat til polsk, litauisk eller tyrkisk.

EU's eget oversættelsessystem, eTranslation, leverer efterhånden ligeledes god kvalitet. Der arbejdes dog stadig på at forbedre eTranslation, især ved at træne systemet på oversatte tekster og oversættelseshukommelser for mindre hyppigt forekommende sprogpar samt på tekster der dækker mange forskellige fagområder. Regelbaseret oversættelse giver ofte de bedste resultater mellem sprog som ikke har mange tilgængelige tekstresurser til træning af statistiske oversættelsessystemer. Fx bruges regelbaserede systemer i dag til samisk og grønlandsk, og der findes også regelbaserede systemer mellem dansk og en række andre sprog, fx portugisisk, catalansk, svensk, norsk, tysk m.fl.[17]

Afdækning af behovet

Både offentlige institutioner og private har brug for at få oversat tekster. Det gælder institutioner som har en høj grad af informationsudveksling over grænserne, fx EU og Norden, fx Udenrigsministeriet eller afdelinger hos TOLD eller SKAT, det gælder institutioner som skal servicere borgere med andre sprog som modersmål, og det gælder virksomheder som handler i udlandet.

Det må forventes at behovet vil stige i de kommende år, ikke mindst i forbindelse med EU's bestræbelser for at fremme det digitale indre marked, hvor automatisk oversættelse pt. udvikles som en del af den digitale infrastruktur der skal gøre det muligt at handle elektronisk og udvikle informationer på tværs af grænserne uden generende tekniske eller sproglige barrierer. EU's seneste tal viser at 90 % af de europæiske forbrugere foretrækker at læse websider på deres eget sprog, og at 42 % aldrig køber på andre sprog end deres eget[18].

Derfor er EU-Kommissionen også særdeles aktiv når det gælder udvikling af automatisk maskinoversættelse og indsamling af flersproglig terminologi, og har i de seneste år sat adskillige projekter i gang for at forbedre sine tilbud, bl.a. stiller EU-kommissionen nu sit eget system eTranslation gratis til rådighed for alle offentlige institutioner i EU.

En vigtig dimension i diskussionen om automatisk oversættelse er spørgsmålet om sikring af fortrolige oplysninger, især persondata, i oversættelsesprocessen. Det er bl.a. EU's vigtigste argument for at tilbyde en intern tjeneste frem for at gøre brug af eksisterende tilbud på nettet. Også den lettiske regering har valgt at udvikle sit eget online-maskinoversættelsessystem mellem lettisk og engelsk for at sikre de informationer som oversættes, mest muligt.

Der findes ikke nogen opgørelse over behovet for oversættelse i danske ministerier, styrelser eller kommuner. Ganske få har interne oversættelsesafdelinger (fx SKAT), langt de fleste udliciterer opgaven, og tendensen

17 https://gramtrans.com/

18 http://ec.europa.eu/commfrontoffice/publicopinion/flash/fl_313_en.pdf

til udlicitering er stigende. Offentlige og private virksomheder kan vælge at få foretaget oversættelse internt i organisationen, fx af en oversættelses- eller kommunikationsafdeling, eller de kan vælge at outsource opgaven til en privat oversætter eller et privat oversætterfirma fx Semantix eller Lionbridge. Der findes også hybride modeller, hvor visse oversættelser outsources til et oversættelsesbureau, mens andre opgaver løses internt afhængigt af teksternes fagområde og fortrolighedsgrad. Udenrigsministeriet har haft en stor intern oversættelsestjeneste, som også varetog oversættelsesopgaver for andre ministerier og havde opbygget en betydelig ekspertise. Afdelingen blev lukket i efteråret 2018, og opgaven er udliciteret til en privat virksomhed. Også andre ministerier og styrelser fx Erhvervsstyrelsen og Kulturstyrelsen får løst deres oversættelsesopgaver af private udbydere.

En undersøgelse foretaget i 15 danske kommuner fordelt over hele landet i 2016 peger på at der på kommunalt niveau ikke arbejdes med en systematisk tilgang til oversættelse. Nogle benytter ansatte der kender det pågældende sprog, men som egentlig er ansat til andre opgaver (23,8 %), andre benytter eksterne oversættere eller oversættelsesvirksomheder (14,3 %), og en mindre del bruger medarbejdere som har sprogarbejde som en af deres hovedopgaver (9,5 %). Tæt på halvdelen af respondenterne, som var udvalgt på grund af deres kendskab til hvordan der arbejdes med sprog i den pågældende kommune, kunne ikke oplyse hvordan den skriftlige kommunikation på andre sprog end dansk varetages i kommunen[19].

Udfordringer

Sprognævnets udredninger, bl.a. i forbindelse med EU's initiativ European Language Ressource Coordination (ELRC), har vist at de fleste offentlige institutioner ikke er bevidste om den samfundsmæssige værdi oversættelseshukommelser repræsenterer.

Manglende viden om sprog og oversættelse og den stigende udlicitering af oversættelsesopgaver i den offentlige sektor forhindrer at der etableres betryggende rutiner omkring det værdifulde oversatte materiale. Det medfører i sidste ende at de offentlige institutioner står svagt i forhold til forhandling af priser, kvalitetssikring af materialet og den samfundsmæssige udnyttelse af de oversatte tekster. Endvidere kan det være vanskeligt at konkurrenceudsætte oversættelsesydelsen når en bestemt virksomhed i løbet af noget tid har opbygget de relevante oversættelseshukommelser. Det gør det vanskeligt for nye udbydere at konkurrere på prisen og risikerer at skabe en uheldig monopolsituation hvis nye udbydere ikke kan få adgang til tidligere oversat materiale.

Der er derfor i den offentlige sektor behov for

- en større bevidsthed om hvilken betydning oversatte tekster har for udvikling af automatisk oversættelse og andre sprogteknologiske applikationer
- en klassifikationsmodel/metadata for oversatte tekster og oversættelseshukommelser som gør det muligt at vurdere om teksterne egner sig til deling, dvs. ikke indeholder persondata eller andre klassificerede oplysninger, og hvilke fagområder og tekstyper de dækker
- klare regler for hvem der har dispositionsretten over de oversatte data i forbindelse med udbud af oversættelsesydelser, fx forpligtelse for private udbydere til at overdrage oversættelseshukommelsen til den offentlige institution eller en offentlig sprogbank
- gode analyser af hvordan det oversatte materiale i øvrigt kan bringes i spil, fx i forbindelse med udvikling af automatisk oversættelse af websider m.m., som resurse for tolkning eller ekstraktion af flersproget terminologi.

2.7. Sprogteknologi for mennesker med behov for kommunikationshjælpemidler

Sprogteknologi har stor betydning for mennesker med behov for kommunikationshjælpemidler og dermed også for at offentlige institutioner kan sikre tilgængelighed til information for alle og skabe lige muligheder for at deltage i samfundslivet. Nogle af de kendte tiltag er It-rygsækken for læsesvage unge og oplæsningsprogrammet Adgang for alle, som kan bruges af ordblinde og svage læsere på alle hjemmesider[20]. Der findes også en række specialiserede hjælpemidler for blinde og svagtseende, fx talegenkendelse og syntese, og her er behovet stort for en større differentiering af udbuddet. Blandt statsinstitutionerne har især NOTA arbejdet med at nyttiggøre taleteknologi, bl.a. ved at indskanne og forsyne lærebøger med oplæsning eller talesyntese.

19 https://dsn.dk/vi-arbejder-ogsa-med/klar-kommunikation/linksamling-1/sprogarbejdet-i-danske-kommuner-2016

20 http://www.adgangforalle.dk/

Sprogteknologi har igennem mange år leveret løsninger for mennesker med behov for alternativ og støttende kommunikation (ASK). På de fleste elektroniske kommunikationshjælpemidler af nyere dato bruges stemmerne Rasmus og Mette. Når ASK-brugere taler i en gruppe, oplever de at alle drenge/mænd har den samme stemme, Rasmus, og at alle kvinder/piger har stemmen Mette. Endvidere har danske børn der benytter talesyntese, ikke adgang til danske børnetalesynteser. Børnetalesynteser findes fx på engelsk.

International Society of Augmentative and Alternative Communication (ISAAC) er en privat interesse- og medlemsorganisation baseret på frivillig arbejdskraft. ISAAC arbejder på at fremme kendskabet til alternativ og supplerende kommunikation (AAC) og opmuntre til, at der bliver taget initiativer til at udvikle området så ALLE mennesker får mulighed for at kommunikere på en værdig og selvstændig måde.

Den danske afdeling af ISAAC, som tæller førende danske eksperter og praktikere inden for ASK[21], peger på følgende fremtidige behov:

- Der er i dag brug for et større udvalg af færdige kvalitetstalesynteser på dansk med mulighed for tilpasninger så borgeren har valgmulighed mellem gode talesynteser
- Danske børn der bruger talesynteser, har brug for danske børnestemmer
- Der er generelt behov for at individualisere talesynteserne så mennesker med behov for alternativ og støttende kommunikation ikke er nødt til at bruge de samme stemmer.

Opfyldelsen af disse behov vil i stort omfang kunne opfyldes af forbedringer af taleteknologien generelt. Der er dog særlige udfordringer i forhold til udvikling af børnetalesynteser, ligesom individualisering af talesynteserne kræver særlige tiltag.

2.8. Sprogteknologi for dansk tegnsprog

Selvom dansk tegnsprog er et helt andet sprog end dansk, har vi valgt at tage det med i udvalgets arbejde da dansk tegnsprog er et sprog der kun bruges i Danmark, og brugere af dansk tegnsprog bør kunne få adgang til at deltage i samfundet og bruge de tilbud der udvikles ved hjælp af sprogteknologi og kunstig intelligens, på lige fod med dansktalende.

Ansvaret for at oplyse om og dokumentere dansk tegnsprog ligger hos Dansk Tegnsprogsråd som sekretariatsbetjenes af Dansk Sprognævn. Dansk tegnsprog er et selvstændigt sprog i Danmark. Ifølge Danske Døves Landsforbund (DDL) findes der ca. 4.000 døve personer i Danmark som har dansk tegnsprog som modersmål (tallet er fra 2014). Det vil sige at lidt under en promille af den danske befolkning har dansk tegnsprog som modersmål. Dertil kommer et antal hørende familiemedlemmer og professionelle der bruger dansk tegnsprog i deres kontakt med døve. Dansk tegnsprog vil være i anvendelse i mange generationer endnu selv om det er et truet sprog, idet antallet af mennesker som bruger dansk tegnsprog er støt faldende fordi mange i dag kan få indopereret cochleare implantater der forbedrer hørelsen.

Der findes allerede en online tegnordbog som udvikles i samarbejde mellem Københavns Professionshøjskole som bl.a. varetager uddannelsen til tegnsprogstolk, og Danske Døves Landsforbund. Der findes endvidere en række apps med ordlister for dansk tegnsprog. Region Nordjylland, Center for døvblindhed og høretab har sammen med Digitaliseringsstyrelsen udviklet tegnprogrammet Adgang med tegn som støtter tegnbrugere med læsevanskeligheder i at læse danske hjemmesider. Værktøjet oversætter udvalgte ord til tegn som vises i korte videoklip uden lyd. Systemet kan bruges som plug-in på offentlige hjemmesider og bruges fx på borger.dk[22]. I de nordiske lande er der etableret et samarbejde mellem tegnsprogsrådene om opbygning af resurser til understøttelse af de nordiske tegnsprog.

For yderligere at kunne understøtte udvikling af kommunikationsmidler til mennesker der bruger dansk tegnsprog, fx metoder til automatisk gengivelse og genkendelse af tegnsprog samt oversættelse mellem dansk tegnsprog og dansk, er der brug for at dansk tegnsprog løbende dokumenteres i form af transskriberede og annoterede videooptagelser.

21 http://www.isaac.dk/om-isaac/

22 http://adgangmedtegn.dk/


Rådhuspladsen

3. Sprogteknologi i Danmark, Norden og Europa

3.1. Det sprogteknologiske landskab i Danmark

Dette afsnit giver et overblik over hvilke former for sprogteknologi der i dag anvendes i Danmark, og hvilke hovedaktører der er involveret i produktion og tilgængeliggørelse af sprogteknologiske resurser.

I en vis forstand begyndte dansk sprogteknologi med WordPerfect. Dette tekstbehandlingsprogram var blandt de tidligste programmer i Danmark som inkorporerede elementer af egentlig sprogteknologi for det danske sprog. WordPerfect's orddelingsalgoritme og stavekontrol blev udviklet af danske professionelle lingvister i midten af 1980erne på blandt andet Institut for Almen og Anvendt Sprogvidenskab (IAAS) og Center for Sprogteknologi (CST) (begge Københavns Universitet) og nåede et teknisk niveau der nærmerede sig datidens state-of-the-art. I de samme år (1985-1995) eksperimenterede flere universtitetsbaserede forskergrupper med talesyntese (især IAAS) og **talegenkendelse** (især Center for Personkommunikation, Aalborg Universitetscenter).

Det første organiserede konsortium til udvikling af **talesyntese** med *industrial strength* blev dannet i midten af 1990'erne og inkluderede TeleDanmark, NOVI Forskerpark, Aalborg Universitetscenter og Københavns Universitet. Erhvervsrettighederne til den producerede kunstige stemme (kaldet Carsten) endte i stort omfang hos TeleDanmark selv om universiteterne beholdt nogle få procent. Efterfølgende, omkring år 2000, opgav TeleDanmark sin satsning på dansk talesyntese, og Carsten blev købt og videreudviklet af Mikroværkstedet A/S (nu MVNordic A/S), som stadig anvender den i læremidler til den danske folkeskole.

Efter år 2000 er talesynteser til det danske sprog blevet udviklet af nogle få danske private og offentlige udviklingsvirksomheder, bl.a. MVNordic, Nota og Dictus, og disse virksomheder arbejder fortsat på at fortsætte udviklingen og på at levere syntesestemmer til markedet. Nogle få europæiske virksomheder (Acapela, SVOX, Ivona, STTS) har uafhængigt udviklet danske synteser, men ingen af disse leverer aktuelt til markedet. Desuden er de fleste af disse firmaer - eller deres taleteknologiske divisioner - i dag blevet opkøbt af større firmaer (fx Amazon, Apple og Nuance), og generelt er ingen af de dansksproglige synteser blevet videreudviklet gennem de sidste 5-10 år, hvorfor ingen af dem kan siges at have kvalitet på internationalt niveau i dag.

Der er adskillige danske syntesestemmer i den frie handel, og skønt kvaliteten ikke altid er i top, er der grund til at forvente at markedets konkurrence og den teknologiske udvikling gradvist vil sørge for de nødvendige forbedringer. Der er derfor mindre grund til at opfatte udviklingen af dansk talesyntese som et offentligt satsningsområde.

Samme tendens mod internationalisering og udviklingsmæssig stagnation er tydelig på området **talegenkendelse**. Skønt det danske sprog for 15-20 år siden havde talegenkendelsesløsninger fra i alt fald tre uafhængige større leverandører, er markedet i dag næsten helt monopoliseret, idet det amerikanske firma Nuance Communications har absorberet en lang række europæiske leverandører såsom Philips Speech Recognition Systems (2008), IBM Speech Group (2009), SVOX (2011) og Loquendo (2011). I flere tilfælde er kun disse firmaers kundekreds blevet vedligeholdt, mens de taleteknologiske produkter selv er trukket ud af markedet.

Generelt har markedsudbuddet lidt af en ringe teknologisk fornyelse, og først de seneste to-tre år er kvaliteten af dansk talegenkendelse blevet mærkbart forbedret som virkning af en voksende konkurrence i markedet. De mange års kvalitetsmæssige stilstand har haft en negativ virkning for taleteknologiens almene omdømme, ikke mindst i kommunerne, som udgør en af landets største professionelle brugergrupper, og dette har i sig selv haft en forsinkende virkning på udbredelsen af taleteknologien, sådan at Danmark i dag må siges at være væsentligt bagud i sammenligning med de øvrige nordeuropæiske lande, ikke mindst Norge, Island og Holland.

Udviklingen af **automatisk oversættelse** i Danmark er dels foregået på Center for Sprogteknologi på Københavns Universitet inden for rammerne af EU-projektet Eurotra i slutningen af 1980'erne og 90'erne og efterfølgende i privat regi via produktet PATRANS omkring 2000, dels på Odense Universitet og efterfølgende i

privat regi via produktet GRAMTRANS omkring 2005. Begge systemer er regelbaserede. Fremkomsten af statistisk oversættelse (i begyndelsen af 2000) og neural oversættelse (omkring 2014) har medført at oversættelsesmarkedet i dag i høj grad er overladt til internationale spillere, med Google som den mest dominerende.

Oversættelseshukommelse arbejder tættere sammen med den menneskelige oversætter og bidrager med tekstforslag i løbet af selve oversættelsesprocessen. Oversættelseshukommelsessystemer er i dag brugt i Danmark på næsten alle oversættelsesbureauer med fokus på fagtekster (fx manualer, brochurer, rapporter, patenttekst) samt i de skrevne medier og på bogforlag, men leveres udelukkende af udenlandske leverandører, fx Memsource, SDL og Wordfast. Oversættelseshukommelser kombineres i dag typisk med maskinoversættelse.

Automatisk tekstanalyse (emneklassifikation, resumering, sentimentanalyse) spiller en rolle i den offentlige administration hvor sagsbehandlingen i kommunerne anvender algoritmer og programmer til emneklassifikation i det såkaldte KLE-system[23]. Desuden anvender store tekstbureauer, såsom Infomedia, algoritmer til tekstanalyse til at monitorere den daglige strøm af nyheder og fagtekster og forsyne deres abonnenter (fx bladhuse og forlag) med relevante udklip og dokumenter. De seneste år har nye anvendelser af automatisk tekstanalyse fået kommerciel interesse, for eksempel ved automatisk kundebetjening og automatisk viderestilling til sagsbehandlere i forsikringsselskaber, banker, teleselskaber og andre virksomheder og organisationer hvor kundebetjening vejer tungt. Automatisk emne- og indholdsklassifikation har et stort udviklingspotentiale i en tid hvor informationsstrømmen er gået over alle bredder. Udviklingen af stærkere programmer til tekstanalyse kræver adgang til store tekstsamlinger (korpusser) egnet til maskinlæring.

Dialogsystemer bruges i dag over hele verden til instruktion af maskiner i værksteder og på byggepladser, assistance til specialister i funktion (chauffører, kirurger, redningspersonale) samt intelligent tekstsøgning over internetbrowser og mobiltelefon. Dialogsystemer har været i almindelig brug i mere end 10 år i USA, og for nylig har Kina erklæret området som en hovedsatsning inden for AI (kunstig intelligens). Industrien herhjemme er først de seneste 6-12 måneder så småt begyndt at eksperimentere med dansksprogede dialogsystemer. Det drejer sig især om små projekter varetaget af det fynske robotmiljø samt demonstrationsprojekter på enkelte universiteter uden direkte kommercielt sigte.

Sidst, men ikke mindst, er en særlig variant af dialogsystemer, nemlig **taleassistenter** som Siri og Google Assistent, begyndt at indtage de danske hjem. Det har til dels øget interessen for sprogteknologi generelt, dels medvirket til at etablere en mere realistisk forventning til taleassistentens intelligens (som er lav) og nytte (som kan være betydelig).

Statslige initiativer til udvikling og distribution af sprogteknologiske værktøjer til tekstanalyse og sprogforståelse

Siden 1990 har der været statslige initiativer til udvikling og distribution af sprogteknologiske værktøjer til tekstanalyse og sprogforståelse i Danmark, typisk i tilknytning til universiteterne. Således blev **Center for Sprogteknologi** i 1991 oprettet som center ved Københavns Universitet med bl.a. dette formål. Siden blev centret omdannet til en sektorforskningsinstitution (1996) og derefter fusioneret med Københavns Universitet (2004) hvorefter centrets primære opgaver blev forskning og uddannelse.

Centret har ansvar for udstilling af sproglige data i forskningsportalen DK-CLARIN, som i dag finansieres af Det Humanistiske Fakultet, Københavns Universitet, og stiller en række sprogteknologiske værktøjer til rådighed i portalen og på centrets hjemmeside, ligesom der fortsat udvikles leksikalske resurser og opmærkede datasæt især med semantisk opmærkning. Center for Sprogteknologi har bl.a. udviklet en sprogteknologisk ordbog (STO) og samarbejder bl.a. med Det Danske Sprog- og Litteraturselskab om at udvikle et dansk ordnet, der beskriver ordenes betydninger og relationer til hinanden, og et dansk framenet, der beskriver de kognitive scenarier som ordene kan indgå i.

Det Dansk Sprog- og Litteraturselskab har gennem mere end 20 år anvendt sprogteknologiske metoder og råder ligeledes over en lang række omfattende sprogresurser af høj kvalitet i form af aktuelle og historiske ordbøger og store tekstsamlinger. Selskabet udvikler også selv sprogteknologiske værktøjer til især tekstanalyse - først og fremmest for at understøtte sine egne ordbogsprojekter. Således var Den Danske Ordbog

23 http://www.kle-online.dk

den første ordbog i Danmark som var baseret på et elektronisk korpus, og selskabet giver på sin hjemmeside og på Ordnet.dk mulighed for at søge i den tekstsamling (Korpus 90 og Korpus 2000) som ligger til grund for ordbogen. Desværre er korpusset belagt med store ophavsretsmæssige begrænsninger og kan derfor ikke deles frit i sin helhed, men Sprog- og Litteraturselskabet er begyndt at stille ordlister og blandede citater fra korpusset til rådighed da disse må deles uden at krænke ophavsretten. Sprog- og Litteraturselskabet har igennem mange år samarbejdet med Center for Sprogteknologi om at udvikle et dansk ordnet og et dansk framenet som ligeledes stilles frit til rådighed. Endvidere har selskabet gjort en stor indsats for at digitalisere en række historiske ordbøger og er involveret i en række projekter om digitalisering af ældre danske litteratur, fx Arkiv for dansk litteratur (ADL). Sprog- og Litteraturselskabets aktiviteter er dels finansieret af en bevilling på finansloven under Kulturministeriet, dels af fondsmidler.

På terminologiområdet blev **DANTERMcentret** oprettet ved Copenhagen Business School (CBS) som et specialiseret center for dansk terminologi og terminologidatabaser i 1998. Centret har ydet rådgivning og vejledning i terminologiarbejde for en lang række private virksomheder og udviklede bl.a. et terminologidatabasesystem, iTERM, som i dag bruges i KMD, Sadolin Albæk, Socialstyrelsen, Kriminalforsorgen, SKAT samt en række udenlandske universiteter. DANTERMcentret udarbejdede i 2011-2014 et forslag til en national termbank i et projekt finansieret af VELUX fonden. I forbindelse med nedprioriteringen af erhvervssprog på CBS blev DANTERMcentret i 2016 splittet op således at databasesystemet udvikles og vedligeholdes af en privat virksomhed, DANTERM Technologies, mens forsknings- og rådgivningsarbejde ikke længere udføres.

Ligeledes på CBS blev **Dansk Center for Anvendt Taleteknologi (DanCAST)** oprettet i 2010 af en kreds af universitetsforskere fra CBS, DTU og Lunds Universitet. Målet for DanCAST var at forske i den samlede proces fra lingvistisk observation (fx fonetiske reduktioner) til udviklingen af praktiske løsninger (fx mere præcise talegenkendere til hospitalerne). DanCAST blev hurtigt et dansk centrum for forskning og udvikling af taleteknologi, ikke mindst på grund af sine hyppige landsdækkende seminarer og kurser for it-ansvarlige i den offentlige sektor og udviklere fra industrien. DanCAST var vært for en række eksternt finansierede forskningsprojekter med bidrag fra kommuner, regioner og virksomheder. Fra 2012-2017 var DanCAST officiel rådgiver for Kommunernes Digitale Fællesskab. DanCAST blev lukket i 2017 i forbindelse med en sparerunde.

Dansk Sprognævn har fulgt udviklingen på sprogteknologiområdet i Danmark med stor interesse igennem de sidste 20 år. Sprognævnet er selv storforbruger af store datamængder i form af tekstkorpusser og udvikler sprogteknologiske værktøjer og indsamler resurser til eget brug. Nævnet har siden begyndelsen af 2018 stillet Retskrivningsordbogen frit til rådighed som en sprogteknologisk resurse for at sikre adgang til den korrekte stavning i sprogteknologiske produkter. Derudover har nævnet en del resurser fra forskellige projekter, fx et fejlstavningskorpus, en stavefejlsliste, en række specialkorpusser, fx elevstile, samt nyordsordbogen Nye ord i dansk som ligesom Retskrivningsordbogen løbende bliver opdateret.

I sommeren 2018 blev Sprognævnets strategi justeret således at den ud over en videreførelse af den udvikling som allerede er igangsat, indebærer et større fokus på de udfordringer som udviklingen inden for sprogteknologi og kunstig intelligens vil udgøre for det danske sprog. Nævnets indgående viden om relationen mellem det talte og det skrevne sprog samt de mangeårige erfaringer med analyse af store datamængder skal nyttiggøres så nævnet kan bidrage til at styrke sprog og kommunikation i det danske samfund i forbindelse med brugen af robotter og kunstig intelligens.

Sprognævnet er sammen med Digitaliseringsstyrelsen og Center for Sprogteknologi de nationale forankringspunkter for EU-projektet ELRC[24] som har til formål at indsamle parallelle tekster fra offentlige institutioner og gøre dem tilgængelige til træning af statistiske og neurale maskinoversættelsessystemer for at forbedre bl.a. EU's automatiske oversættelsessystem eTranslation som stilles gratis til rådighed for offentlige institutioner i hele Europa.

På det seneste har også de mere datalogisk og teknologisk orienterede miljøer i forbindelse med udviklingen inden for kunstig intelligens vist interesse for sprogteknologi og resurser på dansk. Således har **Alexandra-Instituttet** taget initiativ til indsamling af danske sprogresurser[25], og der er lignende initiativer i gang på andre universiteter. Danske Fonde bevilger ligeledes midler til indsamling af resurser og etablering af

24 http://www.lr-coordination.eu/

25 https://alexandra.dk/dk/aktuelt/nyheder/2018/kunstig-intelligens-med-danske-algoritmer

forskningsinfrastruktur, fx har Carlsbergfondet i 2018 bevilget 1,9 mio. kr. til Det Danske Sprog- og Litteraturselskab til digital udgivelse og sprogteknologisk anvendelse af Den Danske Begrebsordbog og 800.000 kr. til IT-Universitetet til udvikling af opmærkede danske sprogresurser.

Det er på den ene side glædeligt at danske ministerier, forskningsinstitutioner og fonde udviser interesse for området og vilje til at understøtte udviklingen af danske sprogresurser. På den anden side er det på høje tid at initiativerne koordineres, og at man sikrer at der udvikles de resurser der er nødvendige for at udviklingen af dansk sprogteknologi for alvor tager fart. Det er altafgørende at resurserne bliver frit tilgængelige og bliver vedligeholdt for at disse investeringer kan få maksimal effekt og komme samfundet som helhed til gode.

3.2. Sprogteknologi i Grønland og Færøerne

Grønland og Færøerne er selvstændige sprogsamfund, og skønt antallet af førstesprogstalende i begge tilfælde er mindre end 1 % af det danske sprogsamfund, har både det grønlandske og færøske samfund overkommet at gennemføre store sprogteknologiske projekter af høj lingvistisk kvalitet. Til begge sprog er der gennem de seneste 8-10 år blevet udviklet stavekontroller, tekstanalyse (fx grammatiske parsere) og oversættelseshukommelser for en række fagområder. I mindre omfang findes også automatiske oversættere, dog kun inden for afgrænsede fagområder. Desuden findes der i dag udmærkede syntesestemmer for både grønlandsk og færøsk. Inden for få år (forudsat at igangværende projekter gennemføres som planlagt) vil også talegenkendelse være tilgængelig for begge sprog. Til forskel fra situationen i Danmark er de nævnte grønlandske og færøske sprogteknologiske komponenter frit tilgængelige for alle borgere.

Rent organisatorisk har Grønland og Færøerne fulgt to forskellige strategier. Det grønlandske sprogsekretariat Oqaasileriffik har oprettet en underafdeling kaldet Oqaaserpassualeriffik (med den direkte oversættelse 'et sted hvor man beskæftiger sig med masser af ord'). Oqaaserpassualeriffik's vigtigste opgave er at planlægge og udvikle sprogteknologiske programmer til Kalaallisut (vestgrønlandsk) inden for tekstanalyse, taleteknologi og automatisk oversættelse. Desuden varetager Oqaaserpassualeriffik indsamling og annotation af en række tekstkorpusser (herunder parallelkorpusser).

På Færøerne har de sprogteknologiske udviklingsprojekter især været baseret på mindre adhoc-konsortier, hvor både virksomheder, offentlige institutioner og det færøske universitet har spillet en aktiv rolle, både som organisatorer og som ansøgere. Færøerne har været aktive i sprogteknologiske projekter støttet af NorForsk, NorFA og andre nordiske forskningsfonde, i samarbejde med Islands Universitet, Copenhagen Business School, Gøteborg Universitet, Trondheim Universitet m.fl. Selv om der således ikke har været en egentlig officiel strategi til støtte for den færøske sprogteknologi, er det alligevel, gennem enkeltpersoners foretagsomhed, lykkedes at fremstille de sprogteknologiske basiskomponenter til færøsk som er nævnt ovenfor.

3.3. Sprogteknologi i Norden og Europa

I de andre nordiske lande har der igennem mange år været fokus på sprogteknologi med forskellige initiativer for de statsbærende sprog finsk, islandsk, norsk, svensk. Norge, Sverige og Finland har haft store satsninger på sprogteknologi og arbejdet med etablering af terminologibaser, annoterede tekstsamlinger, forskningsprogrammer om sprogteknologi m.m. Initiativerne har som i Danmark hidtil ligget i enkeltinstitutioners regi, fx TNC – terminologicentralen (en selvstændig institution) og Språkbanken (en institution ved Göteborgs Universitet) i Sverige. Finland har ligeledes arbejdet med sprogresurseprogrammer, fx FINCLARIN m.fl. I Norge har ansvaret for fx tekst- og taleresurser været placeret i det norske nationalbibliotek. I Island har der været en lang række projekter og programmer med det formål at oprette tekstresurser og især ordbogsresurser for islandsk.

Også for samfundsbærende sprog som grønlandsk og samisk har der været investeret betydelige summer til udvikling af ordbaser og værktøjer til tekstanalyse og automatisk oversættelse. Arbejdet med finsk, samisk og grønlandsk har i høj grad profiteret af at der på grund af sprogenes typologiske fællestræk har kunnet etableres et samarbejde om udvikling af værktøjer og metoder.

Der er en lang tradition for samarbejde blandt nordiske forskningsinstitutioner på det sprogteknologiske område. Arbejdet har bl.a. været støttet af de nordiske landes forskningsråd og af Nordisk Ministerråd gennem NordPlus-programmet. De nordiske sprognævn har igennem en lang årrække i regi af det nordiske samarbejde under Nordisk Ministerråd via arbejdsgruppen ASTIN (Arbejdsgruppen for Sprogteknologi i Norden)

sørget for at følge udviklingen på området. Med jævne mellemrum har gruppen bragt aktører fra forskning, offentlige virksomheder, erhvervsliv og sprogmiljøer sammen for at stimulere vidensudveksling og samarbejde på tværs af landene og på tværs af miljøerne.

Den nyeste udvikling i Norge og Sverige er en tendens til større statslig koordinering af de sprogteknologiske initiativer og resurser i nationale infrastrukturer. Begrundelsen herfor er især effektivitet og synergi ved en samlet tilgang til sprog og kunstig intelligens.

3.3.1. Finland

Finland er et af de få europæiske lande der allerede har en vedtagen strategi for kunstig intelligens. Også her spiller sprogteknologi en rolle, bl.a. i forhold til planerne om at offentlige institutioner skal kunne servicere alle borgere på deres eget sprog ved hjælp af kunstig intelligens og oversættelsesteknologi.
I analysen af Finlands styrker og svagheder i forhold til kunstig intelligens springer et punkt i øjnene som også rammer situationen for sprogteknologi i Danmark[26].

> *The lack of an economy of scale in Finland's operating environment is an unparalleled challenge. We have distributed our resources to separate small projects and no clear focus-point choices or economies of scale have been achieved. This leads us to unintentionally underperforming.*
>
> *Fra: Finland's Age of Artificial Intelligence Turning Finland into a leading country in the application of artificial intelligence. Objective and recommendations for measures. Publications of the Ministry of Economic Affairs and Employment Ministry 47/2017.*

Citatet understreger at en koordinering af indsatsen og en tydelig placering af det overordnede ansvar er af afgørende betydning for indsatsen for sprogteknologi og kunstig intelligens i Danmark.

Finland har endvidere en interessant tilgang til kunstig intelligens som klart også bør overvejes som en strategi for Danmark, nemlig en satsning på at gøre ledere og medarbejdere i virksomheder og offentlige institutioner samt borgerne generelt mere fortrolige med kunstig intelligens. Til dette formål har man udviklet et lettilgængeligt online-kursus i kunstig intelligens "Elements of AI" som er tilgængelig for alle, og som også foreligger på engelsk[27].

3.3.2. Island

Også i Island er der i de seneste år blevet arbejdet politisk og strategisk med sprogteknologi for de islandske sprog. På grund af de relativt få mennesker der taler islandsk, ca. 360.000, er det meget lidt attraktivt for udenlandske producenter at tilbyde islandsk i deres sprogteknologiske produkter. Da Island har en lang tradition for at værne om det islandske sprog, er der fra politisk hold stor opmærksomhed på den teknologiske udvikling, og der er derfor allerede blevet investeret i opbygning af bl.a. ordbaser og korpusser for islandsk igennem en årrække.

Island har vedtaget en strategi for udvikling af sprogteknologi for islandsk[28] som er ved at blive implementeret. Der findes både et strategisk forskningsprogram for islandsk sprogteknologi[29] og en islandsk sprogteknologifond der har til opgave at finansiere projekter som udvikler sprogteknologi for islandsk[30]. Der sigtes især mod udvikling af talegenkendelse, talesyntese, maskinoversættelse, stavekontrol og sprogteknologiske værktøjer som taggere og parsere samt tilgængeliggørelse af relevante datasæt i form af leksikalske data og tekst- og talekorpusser.

26 http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf

27 https://www.elementsofai.com/

28 https://www.stjornarradid.is/lisalib/getfile.aspx?itemid=56f6368e-54f0-11e7-941a-005056bc530c

29 https://en.rannis.is/funding/research/the-strategic-research-and-development-programme-for-language-technology/

30 https://en.rannis.is/funding/research/icelandic-language-technology-fund/

Den islandske sprogteknologifond understøtter følgende formål:

1. Projekter som udvikler nye metoder inden for sprogteknologi eller tilpasning af kendte metoder til skrevet eller talt islandsk
2. Projekter som udvikler nye værktøjer for skrevet eller talt islandsk
3. Projekter som sigter mod at tilpasse og anvende sprogteknologiske værktøjer for islandsk i bestemte fagområder eller i nye anvendelsesmiljøer
4. Projekter som sigter mod at skabe den nødvendige infrastruktur for islandsk sprogteknologi herunder udvikling og vedligeholdelse af sproglige databaser så som leksikalske databaser, tekst- og lydkorpusser, mv. for islandsk.

Alt i alt er der afsat et beløb svarende til 70 mio. DKK til projekter under sprogteknologifonden.

Det islandske forskningsprogram for sprogteknologi ligger under det islandske forskningsråd og følger rådets videnskabelige procedurer for udvælgelse af de ansøgte forskningsprojekter, dvs. med evaluering og prioritering via internationalt anerkendte uvildige eksperter.

I uddannelsessektoren har man netop genetableret en masteruddannelse i sprogteknolog i et samarbejde mellem Háskóli Íslands og Háskólinn í Reykjavík.

3.3.3. Letland

Letland har for længst erkendt vigtigheden af en aktiv politik indenfor sprog og sprogteknologi og er i disse år et ganske aktivt land i det sprogteknologiske felt[31]. Landet har haft flere nationale sprogteknologiske programmer de senere år og har ligeledes været med i adskillige EU-projekter, bl.a. finansieret af EU Structural Funds Programme og EU's FP7. Med støtte fra EU og i samarbejde med private aktører er der oprettet et IT-kompetencecenter, der bl.a. har til opgave at skabe innovative sprogteknologiske løsninger og stimulere det sprogteknologiske marked.

En statusopgørelse fra 2014 viser at de følgende resurser og værktøjer allerede dengang var tilgængelige for lettisk.

- **Sprogteknologiske grundresurser**: Lettiske tekstkorpusser, en lettisk træbank (der har været brugt til at udvikle en lettisk dependensparser), en valensdatabase og en samlet leksikalsk resurse, "Explanatory Dictionary", med indgange fra 225 forskellige lettiske ordbøger fra forskellige tider og domæner
- **Sprogforståelse og informationsekstraktion**: Et lettisk framenet samt en frame-semantisk parser, der bl.a. er anvendt sammen med en navnegenkender og en pronomenfortolker til at skabe et stort informationsekstraktionssystem til det nationale avistekstarkiv
- **Stave- og grammatikkontrol**: Et fejlkorpus til brug for udvikling af stavekontroller, flere forskellige ordklassetaggere og syntaktiske parsere (hvoraf en er udviklet til brug i en lettisk stavekontrol)
- **Maskinoversættelse**: Flersproglige tekstkorpusser, en maskinoversætter der overgår Google Oversæt både i automatiske og menneskelige evalueringer, samt en cloudbaseret terminologiservice (TaaS – Terminology as a Service), der kan integreres i maskinoversættelsessystemer
- **Taleassistenter**: Et talekorpus, der bl.a. er blevet brugt til at udvikle en "smal", lettisktalende virtuel agent (implementeret som en app) der kan omregne valutaer – ekstra relevant i 2014, da Letland det år gik over til euroen.

Senest er man gået i gang med EU-projektet "Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian", som har et budget på i alt 658.000 euro (ca. 5 mio. kr.)[32]. Her skal der opmærkes et korpus med 10-15.000 sætninger både syntaktisk og semantisk med henblik på at skabe en fuldt opmærket resurse der kan bruges til at udvikle en komplet datadrevet sprogteknologisk værktøjskæde for lettisk[33].

31 SKADIŅA, Inguna, et al. Language Resources and Technology in Latvia (2010-2014). In: Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014. IOS Press, 2014. p. 227.

32 Se dokumentet "Letland.EU-finansiering.pdf" i G:\dsn\0000. Arkiv\A 98. Sprogteknologi\Sprogteknologisk udvalg\Rapporten\Afsnit til integration\Om Letland.

33 Se https://github.com/LUMII-AILab/FullStack.

3.3.4. Nederlandene

I Nederlandene har man igennem flere år arbejdet med en strategi for udvikling af en åben, dvs. offentligt tilgængelig, værktøjskasse med basale sprogteknologiske komponenter. De nederlandske og belgiske myndigheder gik sammen om at afsætte sprogteknologiske forsknings- og udviklingsmidler via forskningsprogrammet STEVIN, som løb fra 2004-2011 med et samlet budget på 11,4 mio. euro (84 mio. kr.). Ønsket var at stimulere virksomheder til at udvikle sprogteknologiske produkter for nederlandsk og flamsk[34]. STEVIN-programmet var en koordineret indsats fra det nederlandske finansministerium, det nederlandske forskningsministerium, det flamske institut for innovation, ministeriet for den flamske region og det flamske forskningsråd. For at sikre det nederlandsk-flamske samarbejde hen over grænserne blev programmet placeret i Nederlandse Taalunie, som er en fælles organisation under begge landes regeringer der har til opgave at styrke det nederlandske sprog.

STEVIN havde 3 hovedformål:

1. at formidle viden om sprogteknologiske resultater og at stimulere markedet for sprogteknologiske produkter
2. at fremme strategisk forskning i sprogteknologi og udvikle resurser som er vigtige, og som mangler
3. at organisere opbevaringen, vedligeholdelsen og distributionen af de sprogteknologiske resurser der udvikles.

Videnformidling foregik primært via en sprogportal, nyhedsbreve, konferencer og seminarer der skulle stimulere samarbejde og projektudvikling mellem forskere, udviklere og virksomheder. Stimuleringen af markedet foregik via en række mindre demonstrationsprojekter som offentlige institutioner, virksomheder og forskningsinstitutioner kunne søge om. Der var afsat ca. 1 mio. euro (ca. 7,5 mio. kr.), og der kunne maksimalt søges om 100.000 euro pr. projekt. Blandt de projekter som blev sat i gang, var et værktøj til søgning af nummerplader via talegenkendelse, en talegrænseflade til en lovsamling, og et dialogsystem der kunne informere borgerne om faciliteterne i deres by. Andre videnformidlingsprojekter var en markedsundersøgelse der skulle afdække behovet for sprogteknologiske hjælpemidler for mennesker med læse- og talevanskeligheder, samt et projekt der skulle øge søgningen til sprogteknologiuddannelserne.

Til at fremme den strategiske forskning i sprogteknologi blev der i alt afsat 8,5 mio. euro (ca. 63 mio kr.) som blev fordelt på 3-4 opslagsrunder i fri konkurrence med et panel af internationale eksperter som bedømmere af projekternes lødighed og relevans. Blandt de projekter som blev udviklet, var et automatisk transskriptionsprogram fra tekst til tale, et korpus for udtale af navne for at styrke automatisk navnegenkendelse, et stort almensprogligt korpus for nederlandsk, et system til identifikation af flerordsudtryk i løbende tekst, et system til online medieanalyse samt en udvidelse af det nederlandske talesprogskorpus til også at omfatte stemmer fra ældre, børn og mennesker som ikke har nederlandsk som modersmål[35].

Det var et krav fra starten at de sprogresurser som blev skabt i forbindelse med det strategiske forskningsprogram, blev gjort frit tilgængelige for alle for yderligere at stimulere udviklingen.

Der blev endvidere under Nederlandse Taalunie, som er ejer af alle de udviklede resurser, skabt et sprogteknologisk agentur som har til opgave at opbevare, vedligeholde og videreudvikle resurserne. Agenturet leverede den infrastruktur som var nødvendig for de forskellige udviklingsprojekter. Derved reduceredes udgifter til udstyr, data, software, licenser, eksperter og andet personale samtidig med at resurserne efterfølgende blev sikret optimal synlighed og tilgængelighed ved at de er samlet et sted. Samtidig sikrer agenturet at de resurser som er indsamlet i offentlige forskningsprojekter, ikke samler støv, men forbliver brugbare, fx ved at sørge for at de løbende opgraderes til nye serverprogrammer og styresystemer. Sidst, men ikke mindst, sørger agenturet også for at håndtere ophavsrets- og GDPR-spørgsmål. Agenturet var og er fortsat involveret i de evaluerings- og forhandlingsprocesser der vedrører projekterne. Alle resurser som skabes med agenturets mellemkomst, er åbne frit tilgængelige resurser og gavner således hele det nederlandsk/flamske sprogsamfund.

34 http://www.efnil.org/documents/conference-publications/thessaloniki-2010/language-languages-and-new-technologies/12-Cucchiarini-Bosch.pdf

35 http://www.lrec-conf.org/proceedings/lrec2006/pdf/254_pdf.pdf

STEVIN var organiseret på følgende måde: En programkomite bestående af repræsentanter for forskning og erhvervslivet havde ansvaret for forskningsmæssige og indholdsmæssige spørgsmål. Et internationalt rådgivende panel bestående af anerkendte eksperter inden for sprogteknologi stod for evalueringen og prioriteringen af de indkomne projektansøgninger. Processen blev styret i et samarbejde mellem organisationer svarende til det danske strategiske forskningsråd og Innovationsfonden. De projektforslag som blev indsendt af forskningsinstitutioner og virksomheder, blev først gennemgået af det rådgivende panel og siden vurderet af programkomiteen. Evalueringskriterierne var kvalitet, innovativ kraft, økonomi, projektets bidrag til den overordnede målsætning, håndering af ophavsret og GDPR, brug af standarder og genbrug og inddragelse af eksisterende resurser. Baseret på programkomiteens indstilling formulerede STEVIN-programmets bestyrelse den endelige anbefaling til Nederlandse Taalunie om hvilke projekter der skulle støttes.

Det lykkedes langt hen ad vejen for STEVIN-programmet at skabe en god udvikling for sprogteknologi for nederlandsk, og programmet har fået international anerkendelse. Det er siden blevet fulgt op af andre programmer. I dag optræder nederlandsk typisk på niveau med større sprog som fransk, tysk og spansk mht. sprogteknologisk dækning, fx i kortlægningen af det sprogteknologiske niveau gennemført af organisationen Multilingual Europe Technology Alliance (META).

META-Net-rapporten for Nederlandene konkluderer meget klart: Kun gennem dedikerede programmer som STEVIN blev det muligt at skabe de sprogteknologiske basiskomponenter der gjorde det mere attraktivt for virksomheder at udvikle produkter og tjenester på nederlandsk[36].

Fra: The Dutch Language in the Digital Age, Meta-NET White Paper Series 2012.

3.3.5. Norge

Også i Norge ses en øget interesse for sprogteknologi. Det norske sprognævn, Språkrådet, har sammen med Nasjonalbiblioteket fået en ekstrabevilling på ca. 10 mill. NOK årligt til at opdatere og videreudvikle den norske nationale sprogbank, Språkbanken. Endvidere er der planer om at etablere en stor national terminologiportal i Bergen.

Det norske teknologiråds seneste digitaliseringsrapport "Kunstig intelligens – muligheter, utfordringer og en Plan for Norge" peger ligeledes på behovet for et større fokus på sprogteknologi[37]. Endvidere har det norske sprognævn i 2018 udgivet rapporten Språk i Norge[38]. Her peger man på at sprogteknologi er relevant for alle samfundssektorer, og derfor er placeringen af ansvaret for sprogteknologien i det norske kulturdepartement ikke hensigtsmæssig og bør erstattes af en sektorovergribende placering. Blandt en lang række andre sprogpolitiske mål anbefales en fornyet satsning på indsamling af og arbejde med sprogdata som afspejler norsk tale og skriftsprog, for at sikre at gode sprogteknologiske produkter bliver tilgængelige på norsk.

Rapporten peger endvidere på at produkter baseret på sprogteknologi får stadig større gennemslagskraft, fx bruger allerede 120.000 nordmænd Google Home eller Amazons Alexa. Derfor er det blevet endnu vigtigere – også i et demokratisk perspektiv – at sørge for at sprogteknologiske produkter bliver tilgængelige på norsk.

Googles chefudvikler, Tilke Judd, citeres i rapporten for at påpege at det ikke nødvendigvis er antallet af brugere af et givet sprog som er den afgørende faktor for om en virksomhed beslutter sig for at udvikle et nyt produkt. Det afgørende er derimod hvor store og gode sprogdata der er tilgængelige for det aktuelle sprog.

36 http://www.meta-net.eu/whitepapers/volumes/dutch

37 https://teknologiradet.no/publication/kunstig-intelligens-norge/

38 https://www.sprakradet.no/globalassets/diverse/sprak-i-norge_web.pdf

Generelt rapporteres der om en øget interesse for sprogteknologi, både fra det norske sprognævn og den norske digitaliseringsstyrelse DIFI. Sidstnævnte har bl.a. ansvaret for EU-initiativet CEF Digital. Der er ansat en person med kompetencer inden for sprogteknologi som skal styrke oplysningsarbejdet i forhold til sprogteknologi og begrebsarbejde over for andre offentlige institutioner.

Norge arbejder endvidere på en ny sproglov hvor sprogteknologi efter al sandsynlighed også kommer til at spille en rolle.

3.3.6. Sverige

I Sverige ses en stigende interesse for sprogteknologi inden for forskning og udvikling. Der satses offentlige midler i stor stil for at sikre svensk i sprogteknologiske sammenhænge. Sverige har igennem længere tid haft fokus på forskningsinfrastruktur inden for sprogteknologi. Således har Göteborgs universitet igennem årene opbygget store tekstsamlinger og værktøjer til håndtering af svenske tekster som til dels ligger offentligt tilgængelige, og dette har ført til at Göteborg er blevet værtsinstitution for den nationale sprogbank. På taleteknologisiden er KTH i Stockholm blevet førende. Der satses både inden for forskning og tilgængeliggørelse af kulturarven.

Det svenske forskningsråd, Vetenskabsrådet, har tildelt den nationale sprogbank og forskningsinfrastrukturen SWECLARIN sammenlagt 210 millioner SEK således at der i de næste 7 år kan udvikles en ny national infrastruktur for forskning i sprogteknologi og sprogvidenskab samt andre områder som bedriver forskning baseret på sproglige data[39]. Endvidere uddeler Vetenskapsrådet fra 2018 hvert år 20 millioner SEK til infrastrukturprojekter som understøtter datadreven forskning i samarbejde med Digisam (Samordningssekretariat for digitalisering, digital bevaring og digital tilgængeliggørelse af kulturarven under det svenske rigsarkiv, Riksantikvarieämbetet)[40]. Wallenbergs stiftelse satser i alt 1,6 milliarder SEK på en grundforskningsindsats inden for kunstig intelligens og kvanteteknologi i de kommende 10 år[41]. Den svenske innovationsstyrelse Vinnova har fået til opgave at udrede potentialet for kunstig intelligens i Sverige som grundlag for satsninger inden for kunstig intelligens for at styrke svensk konkurrenceevne[42]. Den svenske regerings tolkeudredningsudvalg har i sin betænkning foreslået at Språkrådet i Sverige sammen med centrale myndigheder udvikler en flersproglig database til tolke- og oversættelsestjenester[43].

3.3.7. Sprogteknologi i EU

Mange europæiske lande har øget deres aktiviteter inden for sprogteknologi i de senere år. En omfattende undersøgelse (European Language Monitor 4 ELM4) foretaget af de europæiske sprognævns samarbejdsorganisation EFNIL i 2017 viser at ca. halvdelen af de EU-lande og associerede lande som deltog i undersøgelsen (i alt 21[44]), har en officiel sprogstrategi, og at endnu flere nemlig 12 ud af 21 har dedikerede støtteprogrammer for sprogteknologi. I 9 af landene rettede strategierne sig imod de officielle hovedsprog, mens to af landene også inkluderede minoritetssprog. Ligeledes angiver 10 lande at de systematisk tilbyder sprogtjenester som fx automatisk oversættelse på offentlige hjemmesider.

Sprogteknologi i EU-Parlamentet

EU-Parlamentet vedtog den 11. september 2018 med et overvældende flertal en resolution om ligebehandling af sprog i den digitale tidsalder.

I oplægget til Parlamentet står der bl.a.:

---

39 https://spraakbanken.gu.se/swe/nyheter/spr%C3%A5kbanken-blir-nationell
http://www.sprakochfolkminnen.se/om-oss/forskning/sprakbanken-sam/nationella-sprakbanken.html

40 http://www.digisam.se/vetenskapsradet-i-samverkan-med-digisam/

41 http://kaw.wallenberg.org/forskning/stor-forskningssatsning-satter-sverige-pa-kartan-inom-ai-och- kvantteknologi?fbclid=IwAR36Mi_NG-ekAj-Densr4c_b9dxtmKhJ06OlFglISLfm3yixWINeHhxX07c8

42 https://www.regeringen.se/pressmeddelanden/2017/12/potentialen-for-ai-i-sverige-ska-kartlaggas/
https://www.regeringen.se/artiklar/2017/12/artificiell-intelligens--en-nyckel-for-att-starka-svensk-konkurrens/?fbclid=IwAR0ts5yLjWFK8F8X-DOwEF2Sc_-cI-OUiVLNB6I5HhymwpHfxfAdqlBHMBgM

43 https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2018/12/sou-201883/

44 De deltagende lande var: Belgien, Bulgarien, Danmark, Estland, Finland, Grækenland, Island, Letland, Litauen, Luxemborg, Nederlandene, Norge, Portugal, Slovakiet, Slovenien, Storbritannien, Sverige, Tjekkiet, Tyskland, Ungarn og Østrig

"Til trods for at sprogteknologier er yderst vigtige elementer i den digitale revolution, er de ikke tilstrækkeligt repræsenteret i de europæiske beslutningstageres dagsorden.
Sprogteknologier bidrager til ligestilling af alle europæiske borgere i deres dagligdag, uanset hvilket sprog de taler. Selv om mindre sprog eller mindretalssprog vil vinde mest ved sprogteknologier, er værktøjerne og ressourcerne til dem ofte knappe og i nogle tilfælde ikke-eksisterende.
Der er reelt en voksende teknologisk kløft mellem store, velfinansierede sprog og de øvrige officielle sprog, sprog med sidestillet officiel status eller ikke-officielle EU-sprog, hvoraf nogle måske allerede er truet af digital udryddelse.
For at slå bro over denne teknologikløft er det nødvendigt med en politik, der fremmer teknologisk udvikling for alle europæiske sprog. Bevarelsen af et sprog og dermed af den kultur, der udvikler sig omkring det, er i høj grad betinget af dets evne til at fungere og være nyttigt i moderne og foranderlige miljøer som den digitale verden.
Kulturel og sproglig mangfoldighed er således tæt forbundet med kapaciteter og ressourcer i den digitale verden."

Fra UDKAST TIL BETÆNKNING om ligebehandling af sprog i en digital tidsalder (2018/2028(INI)). EU-Parlamentet.

Betænkningen foreslår en række politiske løsninger der vil kunne give større sproglig ligestilling i Europa gennem brug af ny teknologi ved at

- forbedre de institutionelle rammer for sprogteknologiske politikker
- skabe nye forskningspolitikker med henblik på at øge anvendelsen af sprogteknologi i Europa
- gøre brug af uddannelsespolitikker til at sikre sproglig ligestilling i fremtiden i den digitale tidsalder
- øge støtten til forbedring af både private virksomheders og offentlige organers udnyttelse af sprogteknologier.

Sprogteknologi på EU-Kommissionens dagsorden

I slutningen af 2018 offentliggjorde Europakommissionen sit forslag til EU's AI-strategi "Koordineret plan for udviklingen og anvendelsen af kunstig intelligens produceret i Europa–2018"[45]. Her konstateres det bl.a. at investeringerne i kunstig intelligens i unionen er lave og fragmenterede i forhold til andre dele af verden som fx USA og Kina. Derfor har EU besluttet "at øge investeringerne til i alt (for den offentlige og den private sektor) mindst 20 mia. EUR i perioden 2018-2020 og øge investeringerne progressivt til 20 mia. EUR om året i det næste årti. Kommissionen øger investeringerne i kunstig intelligens under rammeprogrammet for forskning og innovation Horisont 2020 til 1,5 mia. EUR i perioden 2018-2020, hvilket er en stigning på 70 % i forhold til 2014-2017".

Kommissionen har endvidere foreslået at der afsættes mindst 1 mia. EUR om året fra Horison Europe og fra programmet for et digitalt Europa. For at maksimere investeringerne og samle vigtige resurser som fx data og tilvejebringe sømløse lovgivningsmæssige rammer er det blevet bestemt at alle medlemsstater skal fastlægge nationale strategier for kunstig intelligens og støtteforanstaltninger. Bl.a. arbejdes der med at indsamle flere sprogresurser:

Kommissionens sprogressourcer, som er anvendt til implementering af AI-baseret automatiseret oversættelse og natursprogsbehandling, er blandt de mest downloadede datasæt på den europæiske dataportal. For at forbedre sådanne tjenester planlægger Kommissionen at stille yderligere 10 mio. EUR til rådighed fra Connecting Europe-faciliteten til indsamling af flere sprogressourcer for sprog, der ikke er så udbredt på internettet.

Fra EU's AI-strategi "Koordineret plan for udviklingen og anvendelsen af kunstig intelligens produceret i Europa–2018"

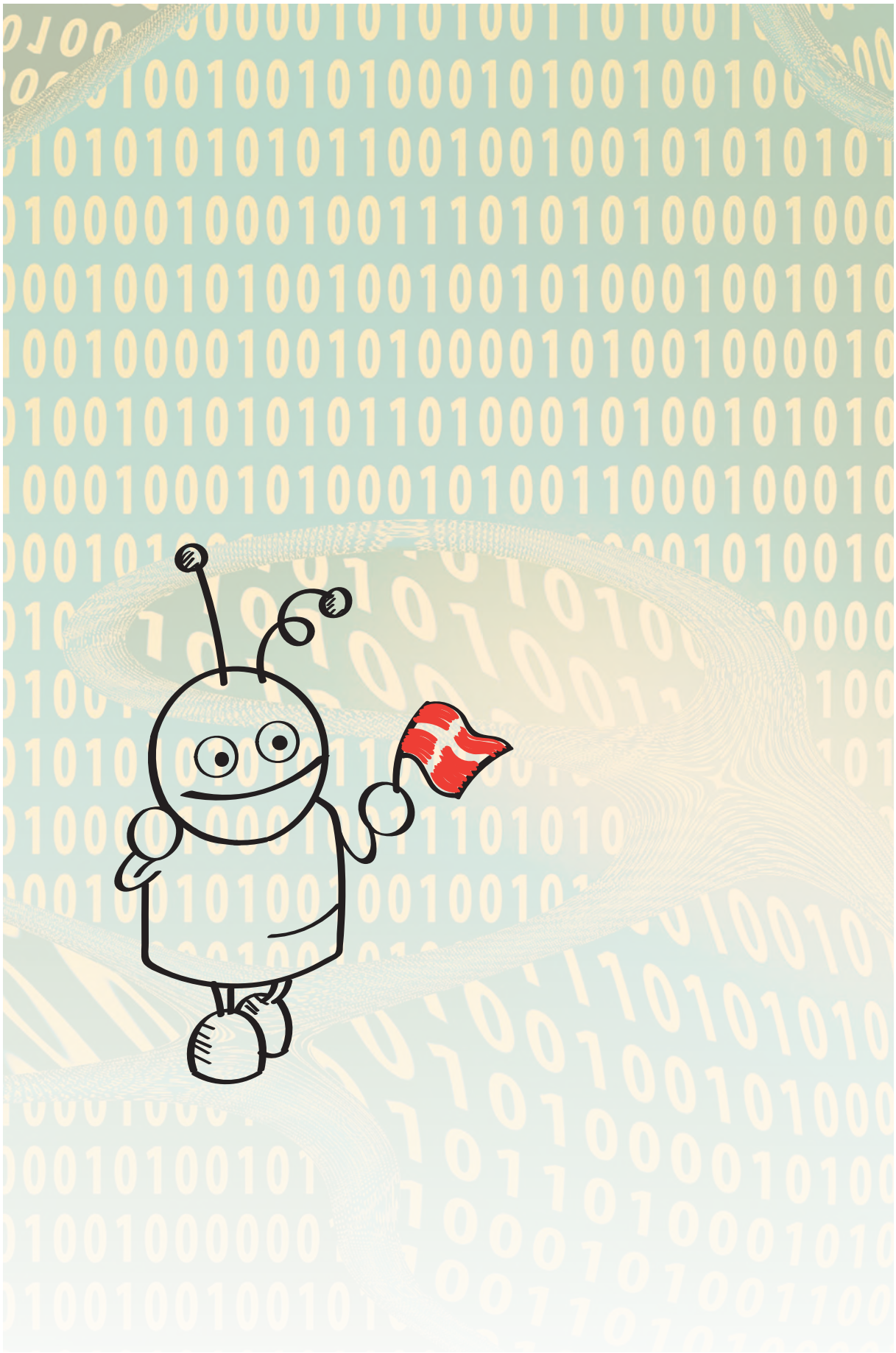
45 https://www.eu.dk/samling/20181/kommissionsforslag/kom(2018)0795/kommissionsforslag/1538436/1987049/index.htm

Kommissionen har allerede iværksat forsknings- og udviklingstiltag til platforme for sikker og kontrolleret deling af ophavsretsbeskyttede data under Horisont 2020, som omfatter industrielle dataområder og persondataområder. På grundlag af Kommissionens meddelelse "Om et fælles europæisk dataområde" er der udgivet et sæt retningslinjer som har til formål at stille en værktøjskasse til rådighed for virksomheder der er dataindehavere, databrugere eller begge dele.

I 2019 vil Kommissionen støtte den næste generation af strategiske digitale industrielle platforme gennem samlende projekter i stor skala med en investering på 50 mio. EUR under Horisont 2020.

Medlemsstaterne opfordres til at forbinde eksisterende og planlagte nationale investeringer i platforme med aktiviteter på EU-plan for at sikre opskalering og interoperabilitet:

I 2019 vil Kommissionen efter planen endvidere tilbyde eTranslation, den AI-baserede automatiske oversættelsestjeneste, der er udviklet under Connecting Europe-faciliteten, til medlemsstaternes offentlige forvaltninger. Kommissionens forslag til Horisont Europa og programmet for et digitalt Europa omhandler investeringer i videreudviklingen af tjenester til natursprogsbehandling og værktøjer til forbedring af flersprogethed i den offentlige sektor.



4. Fremtidens sprogteknologi i et dansk perspektiv

Hvilke danske sprogteknologier er tilstrækkeligt modne til at kunne markedsføres i dag? Hvilke nærmer sig? Og hvad må vi formentlig vente på i mere end ti år? Det kan man få et indtryk af ved at sammenligne med den globale udvikling. Hvert år udgiver det toneangivende amerikanske analysebureau Gartner[46] en graf over tidens nyeste teknologier kaldet the *hype cycle*[47].

Grafen bygger på en antagelse om at alle nye teknologier går gennem de samme stadier i offentlighedens bevidsthed. Først vokser forventningerne voldsomt op til et maksimum der efterfølges af dyb skuffelse over manglende reelle succeser. Derefter opbygges en realistisk forventning, hvorefter det stabile marked opstår. Gartner's hype curve er et af de hyppigst citerede mål for markedets tilegnelse af nye teknologier.

Ifølge Gartner nåede *text-to-speech* (talesyntese) allerede i 2002[48] det modne stadium, mens *text analytics* (automatisk tekstanalyse) og *speech recognition* (talegenkendelse) fulgte efter omkring 2012[49].

Gartners hype curve for 2018 forudser at *virtual assistants* (hjælpefunktioner på internet der bruger naturligt sprog) når frem til det produktive niveau om 2-5 år, mens *conversational AI platforms* kommer omkring 5 år senere (frit talende og tænkende assistenter)[50].

I forhold til Gartners analyser af det internationale marked med USA og Kina som de vigtigste trendsettere er dansk sprogteknologi forsinket med 5-10 år. Dansk talesyntese blev først en handelsvare fra omkring 2010 (på nær synteser til mennesker med funktionsnedsættelser og til begrænsede funktioner i fx GPS og læremidler). Avanceret tekstanalyse og talegenkendelse er stadig ikke tilgængelig i bredt markedsførte produkter og anvendes kun i større skala inden for afgrænsede fagområder (fx sagsbehandling, artikelresumering, oplysningstjenester og diktering).

Danmark er et yderst it-parat samfund og kan relativt let indhente det forsømte. Vi er allerede førende inden for en lang række AI-orienterede teknologier, ikke mindst maskinlæring og robotteknologi. Mange AI-baserede produkter er målrettet mod specialister og kræver betydelig træning at anvende. Ved at kombinere AI med sprogteknologi kan man bringe en lang række intelligente teknologier ind i almindelige menneskers liv.

46 https://www.gartner.com

47 https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

48 https://www.gartner.com/doc/358842/gartners--hype-cycle-emerging

49 https://www.gartner.com/doc/2100915/hype-cycle-emerging-technologies-

50 https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/


Hej
你好

5. Udfordringer for udviklingen af dansk sprogteknologi

Det er relativt kostbart at udvikle sprogteknologi til et lille sprogsamfund. Det danske sprog er ikke simplere i sin struktur eller lettere at behandle teknologisk end fx tysk eller engelsk (i nogle tilfælde snarere tværtimod), men der er færre om at dele udgiften. På den positive side er det ofte lettere at skabe forståelse for nødvendigheden af en national sprogstrategi i et mindre samfund. Som gennemgået i kapitel 3 har bl.a. Norge, Sverige, Nederlandene og Letland kontinuerligt haft store, offentligt støttede sprogteknologiske programmer gennem de seneste årtier som følge af en officiel strategi om at styrke sprogene i alle deres it-relaterede funktioner. I de seneste år har selv sproglige mikrosamfund som Færøerne og Grønland koordineret deres sprogteknologiske tiltag. Danmark har ikke prioriteret en national strategi for sprogteknologi, på nær to mindre udviklingsprojekter i 1998-2002 (talesyntese) og 2003-2006 (talegenkendelse) som ikke udviklede alment tilgængelige basisresurser.

De følgende afsnit gennemgår nogle af de hyppigst nævnte faktorer som i dag virker hæmmende på udviklingen af dansk sprogteknologi: sproglige, teknologiske og kompetencemæssige.

5.1. Sproglige og kulturelle udfordringer

Det danske sprog er, i de fleste henseender, et typisk nordgermansk sprog, i familie med islandsk, færøsk og især svensk og norsk. Dette gælder ikke mindst vores ordforråd, sætningsgrammatik, bøjningssystemer, retskrivning og størstedelen af vores faste vendinger.

På et enkelt område adskiller det danske sprog sig dog markant fra den øvrige nordiske familie: udtalen. Af historiske grunde (som blandt andet har at gøre med indvandringen fra de nordtyske hansastæder i det 15.-17. århundrede og den deraf følgende afsmitning fra nedertysk fonetik) rummer vores sprogs lydside en meget stor mængde forskellige vokalkvaliteter. Hvor norsk og svensk har knapt 20 fuldvokaler (korte og lange), har dansk næsten 40 (stødvokalerne medregnet). Norsk har fx kun én a-kvalitet, mens dansk har tre (sammenlign fx *(en) bane*, *(at) bande* og *barne-(vogn)*). Samtidig er vores udtalevaner i høj grad præget af fonetiske reduktioner der af udlændinge ofte bliver oplevet som 'mumleri'. Tilsammen gør vokalrigdommen og reduktionerne det danske sprog vanskeligt at arbejde med for taleteknologien. En talegenkender bygger på en antagelse om at hvert ord i en sætning kan adskilles klart fra sine naboord og hver lyd fra sin nabolyd, men dette modsiges i høj grad i dansk udtale. Når man i almindeligt taletempo siger … *og I er jo af en anden oprindelse* …, er de første seks ord tilbøjelige til at flyde sammen i én forlænget stavelse. Her giver dagens talegenkendere fortabt.

Der findes i dag ingen udtaleordbog som gør rede for de danske reduktioner. Af samme grund findes der i dag ingen talegenkender som kan analysere danske sætninger med serier af reduktioner som i eksemplet herover (kan efterprøves med Siri eller Google Assisten). En alment tilgængelig sprogteknologisk ordbog over dansk udtalevariation vil derfor betyde meget for videreudviklingen af dansk talegenkendelse.

Det danske sprog giver også udfordringer i forhold til morfologi og grammatik. Når man på dansk typisk kan bøje et substantiv på 8 forskellige måder (*form, forms, formen, formens, former, formers, formerne, formernes*) kan en statistisk behandling af en tekst der ser bort fra dette, ligefrem give misvisende resultater. Derfor kan lemmatiseringsprogrammer og andre programmer der kan håndtere det danske bøjningssystem, gøre en markant forskel. Også sammensatte ord kan volde store problemer. Dansk har let ved at danne lange sammensætninger såsom *medarbejdertilfredshedsundersøgelse* og *forbrugerombudsmandsinstitution* og også dette kan påvirke statistiske og neurale systemer som bl.a. bygger på ordenes frekvens. Nydannede ord er ligeledes vanskelige for stavekontroller, tekstanalysesystemer og oversættelsesprogrammer – og de påvirker pålideligheden i både de statistiske og neurale systemer kraftigt. Selv om disse udfordringer er fælles for de germanske sprog (på nær engelsk) og i nogen grad kan tackles med kendte algoritmer, er der ikke desto mindre et stort behov for en alment tilgængelig database som bl.a. også indeholder regler for orddannelse og algoritmer til morfologisk analyse.

Specielt for dansk er også den omfattende brug af partikler og de sproglige nuanceringer som det medfører, fx er *at skrive op* ikke det samme som *at opskrive, at tage fra* ikke det samme som *at fratage,* og man kan *uddanne nogen,* men man kan ikke *danne nogen ud.* Også i denne sammenhæng kan adgang til avancerede sprogteknologiske ordbøger være en vigtig hjælp til systemernes sprogforståelse og til korrekt generering af fx robotternes ytringer.

Sidst men ikke mindst udgør sproglige traditioner og sprogets historiske og kulturelle forankring en udfordring. Når en robot svarer på dansk, skal den kunne ramme den sproglige norm der gælder for danske samtaler. Allerede små afvigelser skaber irritation hos brugerne. Oversættelse af et system fra fx den engelske kulturkreds til dansk kan give store forståelsesproblemer idet der ikke tages højde for den virksomheds- og samarbejdskultur der hersker i Danmark. Der er især behov for særlig opmærksomhed på anvendelsen af termer idet de afspejler sammenhænge og strukturer som på mange samfundsområder, fx inden for jura, uddannelse og arbejdsmarked, er meget forskellige fra land til land - selv inden for lande som har fælles historiske rødder som landene i Norden.

5.2. Teknologiske udfordringer

De mest udbredte sprogteknologier bygger på maskinlæring og kræver adgang til store mængder af sproglige data (korpusser). Disse anvendes til at optræne programmer til at genkende, oversætte, udtale og på anden måde behandle løbende tekst og tale. Til de fleste typer maskinlæring kræves sprogdata beriget med information om bøjning, grammatik, udtale, betydning, fagligt emne, brugssituation og andre oplysninger (såkaldte metadata). Disse typer af oplysninger kræver ofte manuel analyse, altså at lingvistiske eksperter beriger sprogdataene med de nævnte typer af information. Med korpusser der typisk tæller millioner eller milliarder af ord, bliver denne opgave meget stor. Derfor er korpusser annoteret med lingvistisk information en mangelvare. Selv om simplere typer af annotation kan genereres automatisk (fx ordklasse og bøjning), efterspørger markedet især annotationer af større værdi (især pålidelig information om udtale og betydning). Jo højere kvalitetskrav der stilles, des mere afhænger processen af lingvistisk ekspertise.

I Norden har der været enkelte tiltag til fremstilling og deling af korpusser af høj lingvistisk kvalitet, med konsortiet Nordisk Språkteknologi (NST) som det vigtigste eksempel. Den norske stat investerede i perioden 1997-2003 over 200 mio. kroner i udviklingen af en række annoterede korpusser til norsk og (i mindre grad) andre sprog, herunder dansk. NST fik - inden konkursen i 2003 - oparbejdet en efter tidens forhold betydelig mængde korpusser og ordbaser til maskinlæring. Disse har hverken kvalitet eller kvantitet til maskinlæring efter nutidens standard. Det siger derfor en del om mangelsituationen i Danmark at flere af NST's efterladte sprogresurser stadig den dag i dag bruges som reference i danske forsknings- og udviklingsprojekter. Der er ikke skabt bedre offentligt tilgængelige træningsmaterialer i Danmark gennem de sidste to årtier.

En lang række firmaer og organisationer har i tidens løb stillet korpusser og ordbaser til rådighed for forskningen, fx Dictus, Mirsk, Lingsoft, STTS, Acapela, Mikroværkstedet og mange andre danske og europæiske virksomheder, samt en række danske kommuner, regioner og medier. Disse materialer har dog kun kunnet bruges i kortere projektperioder og under fortrolighedsbetingelser, og ingen er i dag tilgængelige for almenheden.

Også til maskinlæring af semantisk information mangler der semantisk opmærkede guldstandarder og korpusser som et afgørende næste skridt til at kunne udvikle tekstanalyse og tekstforståelse. På grund af de høje kvalitetskrav er opmærkningsarbejdet bekosteligt, og det kræver ydermere modersmålstalende danskere. Derfor kan opgaven vanskeligt løftes af den enkelte virksomhed i forbindelse med produktudviklingen, hvorimod adgang til en sådan resurse vil kunne drive udviklingen af tekstforståelse hurtigere fremad.

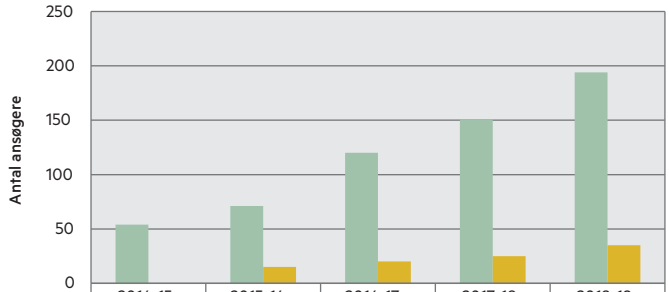
5.3. Kompetencemæssige udfordringer

I 1990'erne havde de danske universiteter to overbygningsuddannelser (cand.mag. og cand.ling.merc.) i datalingvistik. Disse uddannelser var forskningsbaserede, men samtidig direkte rettet mod udvikling af sprogteknologi og med særligt fokus på det danske sprog. Mange af de uddannede kandidater er i dag udviklere og ansatte i danske virksomheder med sprogteknologisk udvikling, heriblandt MVNordic, Gyldendal, PDC, Dictus, Mirsk, IBM Denmark, KMD, Infomedia, Ankiro, Textware, Languagelens, Homunculus, Wizkids og mange flere.

Siden lukningen af den sidste uddannelse i datalingvistik har Danmark manglet nyuddannede sprogteknologer med ekspertviden om dansk sprog. Mange virksomheder har svært ved at finde kvalificerede sprogteknologer med lingvistisk kompetence. Ingen forudser en snarlig løsning på problemet. Københavns Universitet har pt. den eneste kandidatuddannelse med sprogteknologisk indhold, nemlig uddannelsen It & Kognition[51], men uddannelsen foregår udelukkende på engelsk, og der arbejdes som regel ikke med danske data. De fleste universiteter tilbyder endvidere i mindre omfang kurser der omhandler (eller strejfer) udviklingen af sprogteknologi, dog først og fremmest fokuseret på engelsk sprog og almensproglige problemstillinger. På CBS udbydes kurser i kunstig intelligens i forhold til analyse af virksomhedsdata og holdningsanalyser (sentiment-analysis) som enkeltfag eller valgfag. Heri indgår en introduktion til automatisk oversættelse, chatbots og sproglige problemstillinger og til dels statistiske analyser af dansk tekst. På Københavns Universitet udbydes på humaniora et valgfag i it og sprog med fokus på dansk.

Der er en reel risiko for at de studerende der pt. uddannes i it og kognition eller tager valgfag og kurser i sprogteknologi på it-uddannelserne med udgangspunkt i data og værktøjer for engelsk, ikke bliver opmærksomme på de særlige udfordringer ved det danske sprog og det danske marked - og dermed uforvarende bidrager til en situation hvor det danske sprog langsomt forsvinder bag udviklerens horisont.

Der er tilsyneladende et stort potentiale for at uddanne flere kandidater med bedre faglig fundering i dansk sprogteknologi, idet det ser ud til at antallet af ansøgere til både kandidat og valgfag er steget kraftigt i løbet af de senere år.



	2014-15	2015-16	2016-17	2017-18	2018-19
MA IT og kognintion	54	71	120	151	194
Valgfag It og sprog	0	15	20	25	35

Søgning til sprogteknologiuddannelser på Københavns Universitet 2014-2019

I studieåret 2018-19 var der således 194 ansøgere til kandidatuddannelsen og 35 ansøgere til valgfag i it og sprog. Begge steder blev der på grund af dimensioneringen imidlertid kun optaget 25-35 studerende.

5.4. Etiske og juridiske udfordringer

De etiske udfordringer som knytter sig til udviklingen og brugen af sprogteknologi, er til dels de samme som knytter sig til udvikling og brug af kunstig intelligens, og kræver de samme hensyn. Det drejer sig om

- hensynet til beskyttelse af personoplysninger
- hensynet til ophavsretten til data
- hensynet til balanceringen af data.

Disse hensyn er ikke mindst vigtige når data skal stilles åbent til rådighed i offentligt regi.

Et stort antal af de eksisterende sprogresurser er ikke indsamlet med fri deling for øje. Tilladelser fra informanter og rettighedshavere er typisk kun givet til helt specifikke formål og skal genforhandles hvis resursen skal kunne deles frit. Det gælder fx LANCHART-talekorpusset på Københavns Univeristet og dele af resurserne i DK-CLARIN, som udelukkende må bruges til forskningsformål. Det gælder ligeledes Det kongelige Bibli-

51 https://studier.ku.dk/kandidat/it-og-kognition/

oteks Mediestream og Netarkivet, som begge indeholder enorme datamængder som kun i meget begrænset omfang kan stilles til rådighed. Der vil derfor være et udpræget behov for juridisk afklaring af hvordan data kan sættes i spil i forhold til sprogteknologi, og behov for juridisk ekspertise når nye datasæt skal indsamles. Endvidere vil der i mange projekter være behov for juridisk assistance i forbindelse med frikøb af eksisterende resurser.

Særlige hensyn skal tages for at undgå at systemer som er udviklet på basis af store tekst- og datamængder, kommer til at afspejle uønskede skævheder i forhold til fx kønsmæssig, etnisk og religiøs ligebehandling af borgere.

Udfordringer i forhold til beskyttelse af personoplysninger

Når der indsamles store tekstmængder til sprogteknologi som indeholder personfølsomme eller personhenførbare oplysninger, skal GDPR-reglerne naturligvis respekteres. Det gælder dels for teksternes indhold, dels for indhentning af samtykke til fx brug af stemmeprøver m.m. Disse regler respekteres bestemt ikke altid, og der findes omfattende datasæt på nettet som ikke lovligt må bruges, men som alligevel løbende bliver brugt til sprogteknologiske formål. Det gælder fx data fra Facebook og Twitter som kan downloades med henblik på at udvikle systemer til at forudsige shitstorme, og det gælder danske tekstsamlinger som er crawlet fra nettet af udenlandske aktører og efterfølgende indgår i udenlandske portaler, fx Sketchengine[52].

Det nemmeste er naturligvis at sørge for at indsamle tekster som ikke indeholder personoplysninger, fx rapporter og information fra offentlige institutioners hjemmesider, eller tekster hvor deltagerne på forhånd direkte eller indirekte har givet tilsagn om at data må bruges, fx åbne byrådsdagsordener, folketingsdebatter m.m. Det er imidlertid også muligt at bruge sprogteknologiske værktøjer til at anonymisere personoplysninger i materialet, fx ved brug af navnegenkendelse mv.

Der bør udvikles metoder til at forsyne tekster med metadata der indikerer om teksten indeholder personoplysninger, hvilket vil gøre det lettere at sammensætte datasæt til træning af algoritmer, og der bør arbejdes med på forhånd at indhente samtykke, fx fra personer som leverer taledata til forskningsprojekter, med henblik på fri anvendelse til sprogteknologiske formål.

Udfordringer i forhold til beskyttelse af ophavsretten

De ophavsretslige bestemmelser gør at store mængder aktuelle tekstdata, som fx avisernes database Infomedia, eller forlagenes lærebøger og skønlitteratur samt brugergenererede data fx på Facebook, ikke må bruges til sprogteknologiske formål, fordi ophavsretsreglerne forbyder kopiering og deling af disse data med mindre det holder sig inden for citatretten, som er særdeles begrænset. I den forbindelse bør man være opmærksom på at sprogteknologiske systemer typisk ikke er interesseret i disse data som de værker de er ifølge ophavsretsloven, men udelukkende i de sproglige enheder de indeholder. Et projekt der bruger et værk til udvikling af sprogteknologi, vil således ikke på noget tidspunkt gengive værket i sin helhed, men blot anvende værket til fx at træne et klassifikationssystem eller et tekstanalysesystem. Værket vil således kun eksistere som en intern repræsentation i systemet, typisk opsplittet til ukendelighed, og ikke kunne genskabes i sin helhed. Det er særdeles vanskeligt at opnå tilladelse til at bruge store dele af det ophavsretsligt beskyttede materiale fordi der ikke er forståelse for disse særlige forhold for sprogteknologi hos rettighedshaverne. Hertil kommer at de rettighedshavere som har denne forståelse, ofte ikke ønsker at dele data af andre grunde, fx fordi de selv har et ønske om at bruge data til sprogteknologiske eller andre formål. Dermed går et enormt datapotentiale tabt for forskerne og udviklerne.

Det vil kræve en ændring af ophavsretsloven at gøre ophavsretsbeskyttede værker tilgængelige til sprogteknologiske formål. En sådan ændring blev i sommeren 2018 vedtaget i Norge hvor pligtafleveringsloven fik tilføjet en bestemmelse som eksplicit giver mulighed for til forskningsformål at fremstille eksemplarer af værker i andre formater end originaleksemplaret til sproglige korpusser[53]:

52 https://www.sketchengine.eu/

53 https://www.regjeringen.no/no/aktuelt/enklare-tilgang-til-samlingane-ved-nasjonalbiblioteket/id2606628/

§ 1-4 nytt andre ledd skal lyde: Nasjonalbiblioteket kan for forskningsformål fremstille eksemplar av åndsverk i sine samlinger, også i andre format enn originaleksemplaret, som grunnlagsmateriale for språklige korpuser. Dette gjelder også for åndsverk som er omfattet av lov 9. juni 1989 nr. 32 om avleveringsplikt for allment tilgjengelege dokument (Forskrift om endringer i forskrift 21. desember 2001 nr. 1563 til åndsverkloven. Fastsatt av Kulturdepartementet 1. juli 2018).

Fra Forskrift om endringer i forskrift 21. desember 2001 nr. 1563 til åndsverkloven. Fastsatt av Kulturdepartementet 1. juli 2018.

Det bør overvejes om det vil være muligt at indføre en tilsvarende bestemmelse i den danske pligtafleveringslov for at fremme udvikling af og forskning i sprogteknologi på dansk.

Udfordringer i forhold til balancering af data

Især når der anvendes statistiske eller neurale træningsmetoder på sproglige data, er risikoen høj for at der kan opstå ubalance i systemerne således at hensynet til fx kønsmæssig, etnisk eller religiøs ligestilling ikke respekteres. Det sker typisk hvis teksterne er sammensat på en sådan måde at de forfordeler bestemte grupper, eller fordi bestemte grupper simpelthen optræder hyppigere i teksterne end andre og dermed skaber en statistisk ubalance i systemet[54].

Der foreligger endnu ikke forskning der kan give sikre fingerpeg om hvordan disse skævheder kan undgås. Derfor bør der udvises stor omhu i udvælgelsen og sammensætningen af data til fx maskinlæringsprojekter[55]. Sprogteknologi vil formentlig kunne anvendes til at screene tekster med henblik på de faktorer der giver ubalance, men det er et endnu relativt uopdyrket forskningsfelt.

54 http://www.dirkhovy.com/portfolio/papers/download/ethics.pdf

55 http://aclweb.org/anthology/W17-1604.pdf


Ulven er
et rovdyr
...

6. Resurser til udvikling af dansk sprogteknologi

Til trods for de mange udfordringer der er forbundet med udvikling af dansk sprogteknologi, har Danmark også en række muligheder og gode forudsætninger for rent faktisk at skabe dansk sprogteknologi i verdensklasse. Først og fremmest er Danmark et af de mest digitaliserede samfund, og det betyder at store mængder af data i form af dokumenter mv. allerede er tilgængelige digitalt. Der er også etableret gode metoder og retningslinjer for en ensartet sprogbrug og for arbejdet med begrebsafklaring i den offentlige sektor. Der findes danske virksomheder, omend kun få, der tilbyder dansk sprogteknologi, og der findes forskningsinstitutioner og forskere, dog ligeledes kun få, der bedriver god forskning på området. Endvidere har initiativerne omkring digitaliseringen af kulturarven medført at også historiske dokumenter og data i vidt omfang er tilgængelige. Endelig eksisterer der allerede nogle sprogresurser som fx ordbøger, tekstsamlinger og opmærket lyd og tekst som med forholdsvis beskedne midler kan gøres tilgængelige for virksomheder og offentlige institutioner.

En af de store udfordringer er at resurserne ikke er samlet ét sted, men er spredt ud på 72 portaler og hjemmesider hvor de ligger i forskellige versioner. Det er derfor et ganske stort arbejde for en virksomhed eller en forsker at finde frem til de aktuelle resurser og danne sig et indtryk af deres indhold og kvalitet.

Et groft estimat over de eksisterende sprogresurser og datasæt som umiddelbart kan identificeres via forskellige portaler og hjemmesider, ses nedenfor. Estimatet skal tages med forbehold, idet det er et stort og omfattende arbejde at kortlægge resurserne og deres indhold præcist, ligesom oversigten ikke siger noget om resursernes alder og kvalitet. Der er heller ikke medtaget alle resurser som potentielt kunne være nyttige til sprogteknologi, fx er Den Danske Encyklopædi, Trap Danmark og Det Kongelige Biblioteks tekstsamlinger ikke medtaget da det ikke er synligt hvor stort et omfang de har, og under hvilke betingelser de vil kunne bruges.

Oversigt over danske sprogkurser							
Indhold/type	**Parallel tekst**	**Tale**	**Tekst**	**Tekst og tale**	**Termi-nologi**	**Ord-net**	**I alt**
Formelle grammatikker			1				1
Sentimentopmærkning			1				1
Stavning/fejlopmærkning			2				2
Lister over stednavne og geodata			2				2
Træbanker			4				4
Semantisk ordbog eller opmærket tekst			2			3	5
Tekstkorpusser (Ældre litteratur)			6				6
Ordbøger	3	2	10		3		18
Værktøj til danske sprogdata	4	4	20		1		29
Tekstsamlinger og korpusser	11	8	11	1	1		32
I alt	**18**	**14**	**59**	**1**	**5**	**3**	**100**

Oversigt over sprogteknologiske resurser og datasæt for dansk. Den fulde oversigt med links til resurserne kan ses på www.sprogtek2018.dk og på dsn.dk.

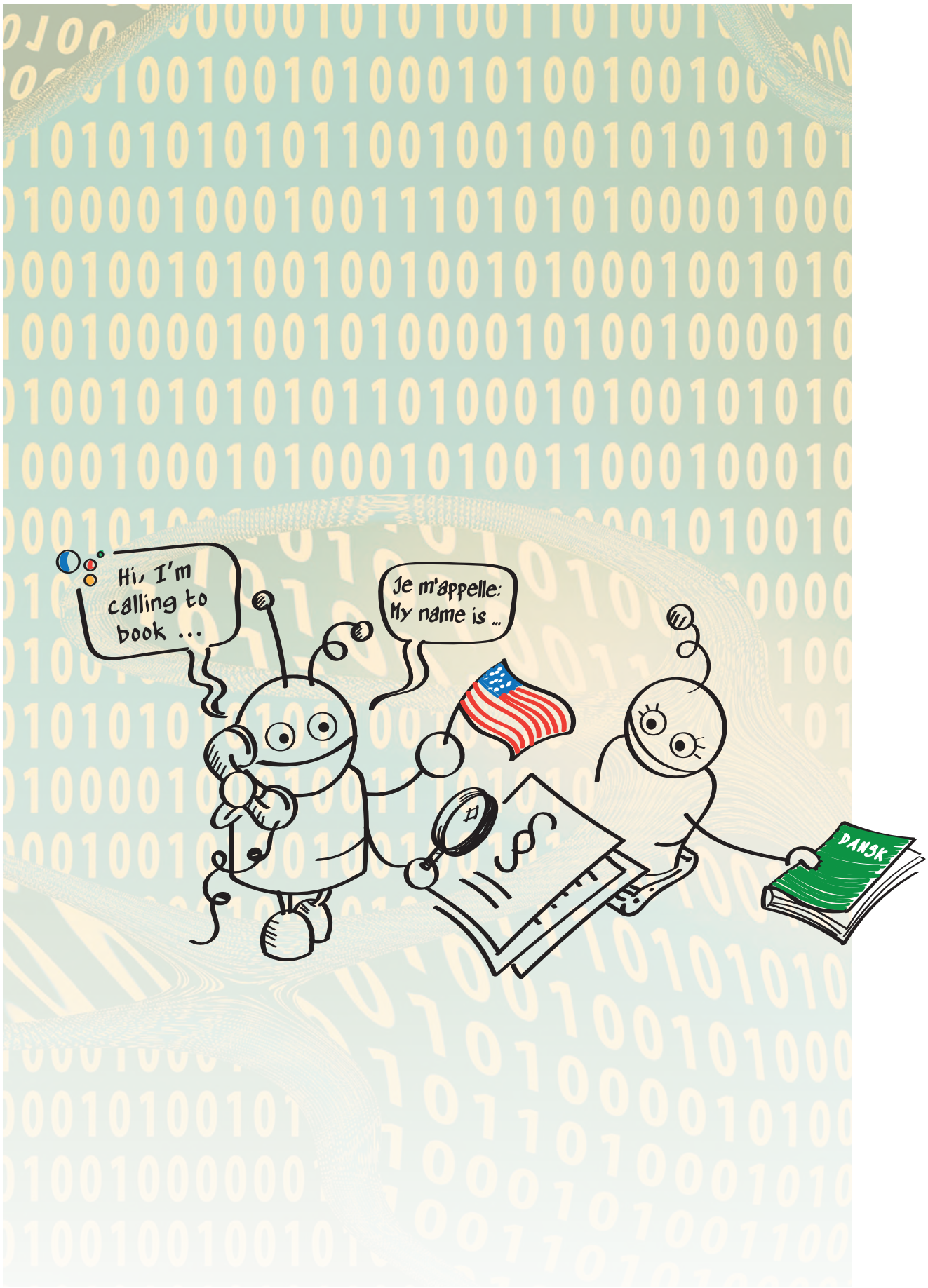
Oversigten viser resurser og datasæt med forskellige typer af indhold, fx grammatisk opmærkning, semantisk opmærkning og tekstkorpusser, og det er angivet hvilke formål de kan bruges til, dvs. parallelle tekster til træning af maskinoversættelse, tekstdata til træning af fx systemer til tekstanalyse og informationsekstraktion, taledata til fremstilling af talegenkendelse og -syntese, term- og begrebsbaser til håndtering af terminologi og ordnet til beskrivelse af relationer mellem ord og begreber. Oversigten siger **ikke** noget om omfanget af de enkelte datasæt og deres kvalitet. Fx kan en resurse som indeholder parallelle tekster, godt bestå af flere mindre delkorpusser. De 32 mindre korpusser som er indsamlet under ELRC med henblik på træning af maskinoversættelse, tæller således kun som ét datasæt. Det samme vil være tilfældet for en række andre resurser.

Det har foreløbigt været muligt at identificere ca. 100 danske sprogresurser, heraf 29 værktøjer som er udviklet til at håndtere dansk. Mere end 3/4 af resurserne er datasæt som består af tekst (ensproget eller parallel), mens resurser som indeholder avanceret opmærkning og taledata, er sparsomme, og en del af dem er forældede. Kun ca. 40 % af resurserne i tabellen er frit tilgængelige. Resten kan man kun få adgang til til forskningsformål eller via en kommerciel licens. En del af dem vil det formentlig være muligt at frikøbe med det formål at gøre dem frit tilgængelige.

Resurseoversigten viser at der allerede findes en del sprogresurser for dansk, men at de ikke dækker tilstrækkelig bredt og ikke umiddelbart er frit tilgængelige. Fra informanterne på udvalgets workshops forlyder det enstemmigt at det er særdeles vanskeligt og tidkrævende at identificere resurserne og vurdere deres anvendelighed, og udvalget vurderer ligeledes at det er en af de største forhindringer for at der udvikles mere dansk sprogteknologi. Det springer især i øjnene at der er meget lidt materiale tilgængeligt for at understøtte udviklingen af dansk tale – mest presserende genkendelse, men også syntese. Endvidere indeholder de kendte resurser stort set ingen dialogdata og semantiske guldstandarder. Terminologidata dækker kun meget få domæner.

Det vil være et større udredningsarbejde i sig selv at gennemgå resurserne i detaljer og at vurdere om de har tilstrækkelig kvalitet, omfang og aktualitet, samt at tage stilling til hvilke der med fordel kan frikøbes for at fremme udviklingen af dansk sprogteknologi, hvilke der bør videreudvikles, og i hvilket omfang der skal tilvejebringes nye resurser.

Oversigten viser at man ikke skal starte fra bar bund, men at der findes en række resurser som man med fordel kan bygge videre på. De enkelte resurser repræsenterer ganske betydelige investeringer som er fortaget af især Kulturministeriet, forskningsministerierne og forskningsrådene hen over en lang årrække. Det skal understreges at disse investeringer om ganske få år vil være spildt hvis ikke resurserne samles og tilgængeliggøres ét sted og løbende vedligeholdes og udvikles.


Hi, I'm calling to book ...
Je m'appelle: My name is ...
DANSK

7. Konklusioner fra sprogteknologiudvalgets workshops

For at kunne afdække behovet for dansk sprogteknologi bedst muligt organiserede sprogudvalgets sekretariat fra maj til december 2018 en række workshops med repræsentanter for virksomheder og organisationer som hhv. bruger, udbyder, udvikler og forsker i sprogteknologi. Derudover blev der gennemført 2 tematiske workshops om hhv. maskinoversættelse og terminologi. Som supplement til workshopperne udsendte sekretariatet spørgeskemaer til workshopdeltagerne. Spørgsmålene skulle belyse nuværende barrierer for brugen af dansk sprogteknologi og indsamle forslag til hvordan disse barrierer kan overvindes.

I det følgende præsenteres de vigtigste konklusioner fra workshopperne.

7.1. Slutbrugere

Deltagerne i slutbrugerworkshoppen var erfarne brugere af talesyntese, talegenkendelse, maskinoversættelse, oversættelseshukommelse og automatisk tekstanalyse i videre forstand (automatisk korrektur, emneklassifikation, sentimentanalyse, tekstanalyse mv.). Vægten blev lagt på de etablerede sprogteknologier da de allernyeste (og fremtidige) applikationsområder, i sagens natur, endnu ikke har erfarne brugere. Der deltog repræsentanter fra kommunale, regionale, statslige og private arbejdspladser, fra de mindste (én ansat) til de største (over 6000).

Spørgeskemaundersøgelse

Forud for workshoppen blev de 35 inviterede bedt om at besvare ti spørgsmål til afklaring af deres faglige profil og professionelle erfaring med dansk sprogteknologi, heriblandt spørgsmål om de teknologier de bruger, hvilke vanskeligheder de støder på, og hvilke teknologier de forventer at komme til at bruge de næste 1-5 år. Spørgeskemaet og besvarelserne er præsenteret i bilag 1.

Det blev tydeligt at systemer til talegenkendelse (især i forbindelse med diktering af tekst), værktøjer til oversættelse (såvel fuldautomatiske som systemer med oversættelseshukommelse og begge dele i kombination) samt systemer til analyse af tekster er de mest anvendte.

Bandt de teknologier man forventede at tage i brug inden for de næste 1-5 år, lå talegenkendelse klart i spidsen efterfulgt af intelligent dialog, livetekstning og sentimentanalyse. Endvidere var deltagerne meget interesserede i at udbrede anvendelsesmulighederne til andre fagområder.

Brugerne savnede især bedre og mere robust talegenkendelse, bedre integration af sprogteknologien med andre systemer og større bredde i fx udvalget af fagområder og stemmer, fx børnestemmer.

Workshop

På workshoppen blev deltagerne bedt om at uddybe deres observationer. Det kom tydeligt frem at centrale lingvistiske og teknologiske udfordringer tydeligvis begrænser sprogteknologiens udbredelse i Danmark.

Blandt de hæmmende faktorer er:

- Der mangler sproglige resurser af tilstrækkelig kvalitet, størrelse, tilgængelighed og genbrugsmuligheder på tværs af brancher og brugergrupper, fx annoterede talesprogsdata, parallelkorpusser for andre sprogpar end dansk-engelsk og kontrollerede emne- og termbaser for en lang række fagområder
- Der opleves store teknologiske udfordringer på grund af de særlige danske sprogtræk. Det gælder fonetiske fænomener som fx stød, tryk og reduktion og leksikalske træk som fx fri kompositumdannelse og faste flerordsforbindelser

- De teknologiske paradigmer som anvendes i mange kommercielle applikationer, er forældede, fx bruges segmentalfonetisk baseret talegenkendelse og n-grambaseret oversættelse uden syntaktisk dybdeanalyse.

Der viste sig også mange praktiske og organisatoriske problemer med indføring af sprogteknologiske værktøjer.

- Der mangler viden hos slutbrugerne (både operatører og ledelse) om hvordan teknologien fungerer, og hvordan virksomheden kan bidrage til at forbedre systemerne
- Der mangler forståelse for at efterbehandling af data ofte er nødvendig, og der afsættes ikke tilstrækkelige resurser til det
- Brugergrænsefladerne er rigide og tvinger brugere til at inddatere efter en fast og arbejdstung rutine
- Der savnes god systemintegration, dvs. bedre integration af sprogteknologiske værktøjer i virksomhedernes samlede workflow, fx sagsbehandlingssystemer, patientjournaler mv.
- Der mangler viden hos slutbrugerne om at virksomhedens sprogdata, fx dokumenter og annotationer, udgør en værdifuld resurse til udvikling af dansk sprogteknologi i et nationalt perspektiv
- Der mangler viden hos slutbrugerne om hvordan de sikrer sig ejendomsretten til virksomhedens sprogdata, fx dokumenter og annotationer, når de indgår aftaler med leverandørerne.

Mange deltagere påpegede at der er en stor risiko forbundet med at indføre sprogteknologi hvis man ikke har tilstrækkelig indsigt i de krav systemerne stiller til brugerne. Virksomhederne fokuserer typisk på den praktiske indføring af teknologien og på de besparelser den kan medføre, men ikke på at uddanne brugerne, som ofte ikke har særlige sproglige forudsætninger for at betjene systemerne. Det kan medføre at brugerne afviser systemet for hurtigt og undgår at bruge det.

"Også i regionen støder jeg ofte på det faktum at viden om sprogteknologi, og ikke mindst udfordringerne med dansk, er meget lille. De fleste har ikke erfaring med området og har derfor, naturligt nok, svært ved at forstå hvor ressourcetungt det er at udvikle sprogteknologi til dansk og at tilpasse indkøbte værktøjer og systemer til domænespecifikt dansk.
Når en teknologi fungerer godt på engelsk eller alment dansk, kan det være umuligt for andre end specialister at vurdere hvad det kræver at få det samme til at virke på dansk inden for et specifikt domæne.
Derfor bør ikke mindst ledelse og politikere rådgives bedre.
I det offentlige kunne det evt. varetages af sprogteknologiske konsulenter, placeret centralt eller decentralt i større instanser, og som samarbejdede på tværs af instanserne med henblik på videndeling, sparring og genbrug." (Deltager i workshop).

7.2. Leverandørperspektivet

I Danmark findes der ca. 20 større leverandører af dansk sprogteknologi, dvs. danske virksomheder eller internationale virksomheder med afdelinger i Danmark med mere end 10 ansatte, samt en lang række mindre virksomheder. Alle firmaer med dansk sprogteknologi i produktporteføljen blev kontaktet, dels med et spørgeskema til individuel besvarelse, dels med en invitation til at deltage i en workshop.

Spørgeskemaundersøgelse

I spørgeskemaet ønskede vi især at få belyst fordelingen af produkter over de gængse sprogteknologier, leverandørernes holdning til samarbejde (med kunder og konkurrenter) samt fremtidens nye muligheder i det danske marked. Spørgeskemaet og besvarelserne er præsenteret i bilag 1.

Workshop

På baggrund af besvarelserne blev de fremmødte informanter inddelt i fire diskussionsgrupper. Hver gruppe blev opfordret til at udarbejde en SWOT-analyse. Resultatet blev derefter præsenteret af hver gruppe i plenum. Forsamlingen udarbejdede derefter en fælles SWOT som en syntese af alle gruppers synspunkter. Den fælles SWOT er resumeret herunder, mens de enkelte gruppers SWOT-analyser er fremlagt i bilag 1.

Styrker	Svagheder
• Branchen er god til at udvikle kreative løsninger i samarbejde med forskere og offentlige institutioner • Der er mulighed for at servicere fagligt snævre brugergrupper pga. en købedygtig offentlig sektor (fx handikapløsninger, undervisningsmidler, velfærdsteknologi) • Stat, regioner og kommuner er aktive medudviklere og villige til at finansiere • Markedet er modent • Der er en række velafprøvede produkter • Markedet er opdelt hvilket giver mindre konkurrence og tæt samarbejde med kunderne • Løsninger udviklet for dansk kan tilpasses til andre sprog.	• Svært at lave en god business case for dansk • For lav sproglig kvalitet i produkter (sammenlignet med engelsksproget teknologi). Udvikling af højere kvalitet er kostbar: mindre virksomheder holder sig tilbage på grund af omkostningerne, større virksomheder finder ikke markedet tilstrækkeligt attraktivt • Der mangler frit tilgængelige sprogresurser af høj kvalitet til udvikling og test af produkter • Det danske sprog giver særlige udfordringer som de internationale teknologier ikke understøtter (fx sammensatte ord, reducerede udtaler og det komplicerede vokalsystem). Der er en tendens til at snævre domæner ind (fx sundhed) • En stor del af finansieringen er offentlig (private investeringer ønskes) • Der er for få spillere i DK, for lidt konkurrence • Uddannelsesgrundlaget er for svagt inden for den lingvistiske kernefaglighed.
Muligheder	**Trusler**
• Delte, åbne databanker (konversationsdata, taledata fra telefoni …) • Mulighed for at bevæge sig i nye retninger, nye sprogteknologier i anvendelse på de domæner der allerede betjenes • Mulighed for at samarbejde om standarder og basale resurser hvilket kan give styrkeposition internationalt • Selv om markedet er lille, indeholder det meget kapital ift. størrelsen • Mulighed for at få støtte fra fx EU • Mulighed for at etablere et uafhængigt, objektivt, offentligt evalueringsudvalg der kan opliste og kvalitetsevaluere eksisterende resurser • Mulighed for at stimulere markedet med shared tasks-konkurrencer • Mindre virksomheders specialviden kan med fordel opsamles af staten efter projektslut • Større fokus på eksport • En sammenhængende terminologibase med dansk udgangspunkt vil få en god start • Oprettelse af en brancheorganisation eller erfa-samarbejde for danske leverandører af sprogteknologi • Mulighed for langt bredere anvendelse af sprogteknologi i den offentlige sektor forudsat at staten kan tilbyde centralt udviklede tjenester til anonymisering, benchmarking, opmærkning (fx af faglige emner) og kvalitetssikring af sprogteknologiske produkter.	• Konkurrence fra udenlandske konkurrenter som tilbyder produkter af ringe (dansk-)sproglig kvalitet, men har bedre muligheder for markedsføring • Svigtende uddannelse af datalingvister • Sprogteknologi fylder for lidt inden for uddannelser i it og kommunikation og på professionshøjskolerne • Ingen dialogsystemer på dansk • Dansk bliver mere og mere negligeret/nedprioriteret af/i de store sprogteknologivirksomheder • Maskinlæring medfører at grundlæggende viden om sprog forsvinder • Datakvaliteten er for lav til god maskinlæring • Der mangler benchmarkingstandarder • De danske forbrugere skal leve med dårlig behandling af det danske sprog i hverdagens it (taleassistenter, telefontjenester, oplysning, bil/GPS osv.) - og mister derved tilliden til dansk sprogteknologi generelt.

7.3. Udviklerperspektivet

55 professionelle udviklere af dansk sprogteknologi som blev identificeret i udredningsarbejdet, blev inviteret til workshop. Næsten alle har ud over professionel erfaring med udviklingsarbejde også andre funktioner inden for marketing, salg, forskning, undervisning og rådgivning.

Spørgeskemaundersøgelse

Spørgeskemaundersøgelsen havde til formål at kortlægge udviklernes specifikke udfordringer i forhold til dansk sprogteknologi. I modsætning til udbyderne var fokus ikke på konkurrence, kunder og markedsføring, men mere på de tekniske og resursemæssige muligheder for at levere et godt sprogteknologisk produkt. Udviklerne blev bl.a. spurgt om hvilke sprogteknologier og sprog de arbejder med.

Sprogteknologier:	**Målsprog:**
Tekstanalyse og tekstklassifikation (5) Taleteknologi (5) Korpus og korpusværktøj (2) Ordbøger (2) Maskinoversættelse (1) Dialogsystemer (1)	Dansk (10) Andre nordiske sprog (7) Engelsk (4) Andre europæiske sprog (4)

Hovedvægten viser sig at ligge på tekstanalyse og taleteknologi, mens maskinoversættelse og dialogsystemer, fx chatbots, er mindre udbredt. Det er endvidere tydeligt at de fleste arbejder med danske eller nordiske sprog, mens kun et fåtal arbejder med engelsk eller andre europæiske sprog. Cirka en fjerdedel af besvarelserne repræsenterer virksomheder uden egen udvikling af sprogteknologi, men med anvendelse af sprogteknologi i andre produkter. Langt den overvejende del af udviklerne arbejder med regelbaserede metoder, cirka halvdelen arbejder med både regelbaserede og statistiske metoder. Cirka halvdelen angav endvidere at de arbejder med fagsprog eller af og til arbejder med fagsprog.

Blandt de mest markante tilbagemeldinger i spørgeskemaundersøgelsen var at kvantiteten og kvaliteten af de resurser som udviklerne har til rådighed, ikke er fremragende bortset fra dem de selv har produceret og forædlet igennem en lang årrække. Denne situation binder typisk virksomhederne til de produkter som de selv har udviklet resurser til, og gør det vanskeligt at udvide til nye produkter, nye fagområder og dermed et større marked.

> *"Vi har prøvet at bruge dansk Wikipedia som grunddata, men det havde vi ikke gode erfaringer med. Vi regner med at det skyldtes at sproget var væsentligt forskelligt fra det data vi analyserer." (Svar i spørgeskema).*

Et andet aspekt er adgang til viden om danske sprogteknologiske værktøjer og resurser. Også her melder udviklerne tilbage at det er svært at følge med udviklingen, at finde frem til hvilke resurser der er tilgængelige, og hvilken kvalitet man kan regne med.

> *"Der findes meget forskning, men det er uoverskueligt at udvælge lige det vi kan bruge. Og når vi finder noget, vi gerne vil bruge, så er det besværligt at tilpasse et GitHub bibliotek, så det passer ind i vores platform til vores data." (Svar i spørgeskema).*

Workshop

På baggrund af spørgeskemaundersøgelsen blev de fremmødte informanter inddelt i fire diskussionsgrupper. Hver gruppe blev bedt om at tage stilling til to scenarier.

1. Hvad skal en samling af sprogteknologiske resurser, fx ordbøger, korpusser, regelsamlinger, tekstarkiver, taleoptagelser m.m., indeholde for at være mest nyttig for jer?

2. Hvilke services skal et offentligt videnscenter for sprogteknologi, fx til rådgivning, efteruddannelse, netværksformidling, kontaktskabelse m.m., tilbyde for at være mest nyttigt for jer?

Gruppernes besvarelser blev gennemgået og samlet i en række overordnede ønsker og anbefalinger.

Ønsker til en dansk sprogbank

- En frit tilgængelig **orddatabase** med en dækning af det almene ordforråd svarende til det man finder i eksisterende danske ordbøger fx Den Danske Ordbog og Retskrivningsordbogen (hhv. 100.000 og 65.000 opslagsord) og desuden dækning af udvalgte fagområder.
  - Orddatabasen skal være forsynet med information om stavning, bøjning, ordklasse, udtalevariation, forekomst, brug, frekvens, betydning osv. knyttet sammen i et applikationsneutralt, åbent og bredt anerkendt format
  - Orddatabasen skal vedligeholdes og udvikles løbende da den ellers hurtigt vil blive forældet og miste sin værdi på grund af sprogets udvikling (nye ord, nye vendinger, nye brugssituationer osv.).
- Et udvalg af annoterede **tekst- og taleresurser** som kan deles frit (i størrelsesorden af milliarder ord) i fælles annotationsstandarder (åbne standarder), herunder:
  - Almensproglige tekster og lydoptagelser
  - Fagsproglige tekster med emneklassifikation. Eksempler: medicin, jura
  - Tekst- og taleresurser af særlig høj kvalitet (med manuelt sikret annotation), der kan fungere som guldstandard til maskinlæring
  - Alle tekst- og taleresurser skal være **fri for rettighedsmæssige bindinger** og skal kunne anvendes til fx maskinlæring.
- En **værktøjskasse** af sprogteknologiske programmer målrettet det danske sprog som kan deles frit, fx standardværktøj til opmærkning og analyse af sproglige data.
  - Der lægges vægt på at værktøjer er platformsneutrale og baseret på åbne dataformater og programmeringssprog
  - Værktøjer kan fx have form af programpakker, API-baserede tjenester og SaaS (software as a service)
  - Alle værktøjer skal være forsynet med manualer udarbejdet efter fælles standarder og udviklet løbende i kontakt med brugere.

Ønsker til et videnscenter for dansk sprogteknologi

- Adgang til **rådgivning** på højt fagligt niveau om sprogteknologiske muligheder og løsninger for især dansk sprog samt om standarder for korpusopmærkning, termbanker etc.
- Adgang til korte behovsstyrede **efteruddannelsesforløb** om konkrete emner
- Adgang til et **kontaktnetværk** og netværksaktiviteter, fx om underleverancer business-to-business, samarbejde mellem offentlige og private virksomheder, forsknings- og udviklingssamarbejde
- Formidling af samarbejde med **uddannelser** inden for dansk sprogteknologi, fx universiteter, professionshøjskoler m.fl.).

7.4. Forskning, undervisning og formidling

Den sidste workshop var rettet mod forskere og undervisere med indsigt i dansk sprogteknologi og de særlige udfordringer som det danske sprog indebærer, samt mod lærere, forskere og formidlere der har inddraget dansk sprogteknologi i egen forskning, formidling og undervisning. For at sikre workshoppens fokus på dansk sprogteknologi undlod vi at invitere forskere og undervisere inden for fx almen NLP (natural language processing), almen AI (artificial intelligence) og andre teknologiske områder som ikke sigter mod udvikling af sprogteknologi for det danske sprog.

Spørgeskemaundersøgelse

Spørgeskemaet blev udsendt til 40 informanter og indeholdt spørgsmål om forventninger til fremtidens arbejdsmarked, mulige anvendelser for sprogteknologi, forskningsbehov m.m. Hele spørgeskemaet er gengivet i bilag 1.

Workshop

På den efterfølgende workshop blev en række af de overordnede problemområder som de tidligere workshops havde afdækket, taget op til diskussion:

1) Efterspørgslen i samfundet efter sprogteknologiske løsninger vokser stærkt i disse år (borgerservice, læremidler, kompenserende sproghjælp, oplysning, betjening af maskiner og robotter m.m.). Hvilke initiativer bør staten tage for at imødekomme efterspørgslen?
2) Dansk erhvervsliv og forskning efterspørger sprogteknologiske resurser (databaser, korpusser, ordbøger, termbanker, redskaber) af høj lingvistisk kvalitet. Hvilke initiativer bør staten tage for at imødekomme efterspørgslen?
3) Der mangler kandidater med gode dansk-lingvistiske kompetencer inden for sprogteknologi, både i den private og offentlige sektor. Hvilke initiativer bør staten tage for at imødekomme behovet?

Diskussion førte frem til følgende generelle konklusioner:

1) Hvordan kan man imødekomme efterspørgslen i samfundet efter sprogteknologiske løsninger

- Der bør arbejdes hen imod mere sproglig viden i de sprogteknologiske løsninger - der mangler dansk sprogteknologi med høj sproglig kvalitet
- Der bør oprettes puljer til at udvikle løsninger også på mindre sprogområder (fx tegnsprog)
- Der bør etableres et center der koordinerer leverandører og aftagere, erhverv og offentlige institutioner osv.
- Virksomheder bør have bedre adgang til konsulentydelser inden for sprogteknologi (opmærkning fx).

2) Hvordan kan man imødekomme efterspørgslen i samfundet efter sprogteknologiske resurser

- Der bør udvikles sprogresurser (opmærkede korpusser og ordbaser) af høj kvalitet lavet af sprogeksperter
- Regelbaserede systemer af høj kvalitet kan bruges som supplement til statistiske og neurale metoder
- Der bør etableres åben og nem adgang til data, gerne vha. webservices
- Resurser skal kunne deles som open source, fri for ophavsret
- Der bør være bedre mulighed for at søge midler til udvikling af nye resurser
- Staten bør arbejde for at fjerne ophavsret på tekstmaterialer
- Der bør oprettes målrettede forskningsprogrammer så forskerne kan søge midler til udvikling af dansk sprogteknologi.

3) Hvordan kan man imødekomme efterspørgslen i samfundet efter sprogteknologisk ekspertise

- Der er brug for sprogteknologer med stærke tværfaglige kompetencer i krydsfeltet mellem lingvistik, datalogi og data science
- Der er brug for en bachelor- og en kandidatuddannelse med fokus på dansk sprogteknologi
- Der er brug for grundkurser i digital humaniora for at brede kendskabet til dansk sprogteknologi mere ud
- Data science bør suppleres med sprogteknologi i nogle fagspecialiseringer, fx på datalogi/ITU
- Der er et stort behov for korte virksomhedsrettede kurser og efteruddannelse.

7.5. Automatisk oversættelse

Workshoppen om automatisk oversættelse fandt sted den 8. oktober 2018 i Dansk Sprognævns lokaler i København. Vært for workshoppen var Dansk Sprognævn og den danske EU-repræsentation idet workshoppen foregik i regi af EU's Connecting Europe Facility (CEF) som bl.a. står bag projektet European Language Ressource Coordination (ELRC)[56]. Et fyldigt referat af workshoppen findes på projektets hjemmeside[57].

I alt 30 deltagere fra danske offentlige og private institutioner, fra fagforeninger og fra ELRC lyttede til danske og internationale foredrag om automatisk oversættelse og diskuterede hvordan det bedst kan sikres at sprog med få millioner sprogbrugere også kan blive oversat automatisk med en høj kvalitet af EU's sprogservice eTranslation og andre. Fra EU's side understregede man især behovet for at fjerne sprogbarrierer i det digitale indre marked og fx gøre det muligt at drive e-handel på alle EU-sprog og at levere offentlige serviceydelser på nettet til alle EU-borgere på hver deres sprog.

Det blev konkluderet at der i store dele af den offentlige sektor stadig mangler bevidsthed om hvor værdifulde tekster og terminologier er for at man kan skabe bedre oversættelser og bidrage til forskning og udvikling af sprogteknologi for dansk. Mange institutioner indser ikke fordelene ved at oprette og vedligeholde tekst- og terminologidatabaser, eller de kan ikke finde de nødvendige resurser til disse opgaver. Nogle arbejder systematisk med terminologi, men ikke med tekstdata. Mange offentlige institutioner har outsourcet deres oversættelsesopgaver til private leverandører uden at træffe de nødvendige foranstaltninger til at bevare, organisere og genbruge de oversatte data til andre formål.

For at EU's oversættelsessystem og en kontinuerlig indsamling og deling af danske tekstdata skal blive en integreret del af arbejdsgangen i danske offentlige institutioner, bør det også være muligt for private leverandører at anvende EU's oversættelsessystem hvis de oversætter EU-relaterede tekster eller løser andre typer af oversættelsesopgaver for offentlige institutioner.

Et offentlig-privat samarbejde om oversættelse og brug af offentlige sprogdata reguleret af fx reglerne for offentlige udbud kunne sandsynligvis forbedre situationen i Danmark betydeligt.

7.6. Terminologi

Workshoppen om terminologi fandt sted den 12.12.2018. Der deltog i alt 28 repræsentanter for danske offentlige institutioner og private virksomheder.

Forud for workshoppen var der udsendt et omfattende spørgeskema til ca. 100 virksomheder og offentlige institutioner. Det blev besvaret af 61 personer, dvs. en svarprocent på lidt over 60. Besvarelserne var meget jævnt fordelt med et lille flertal af offentlige institutioner på 56 %. Resten var private virksomheder eller organisationer.

Svarene kom bl.a. fra

- Myndigheder (styrelser, ministerier, Folketinget), Regioner, Det Europæiske Regionsudvalg, Danmarks Statistik, Nationalbanken
- Virksomheder (KMD, banker, pensionsvirksomheder, produktionsvirksomheder, konsulentvirksomheder, it-udviklere, oversætterbureauer)
- Højere læreanstalter (institutter/centre vedr. sprog og kommunikation)
- A-kasser
- Sprog- og kommunikationskonsulenter
- Tolkeforeninger og -virksomheder, netværk af sprogmedarbejdere
- Europa-Parlamentet, Europa-Kommissionen, Oversættelsescentret for Den Europæiske Unions Organer

56 http://www.lr-coordination.eu/

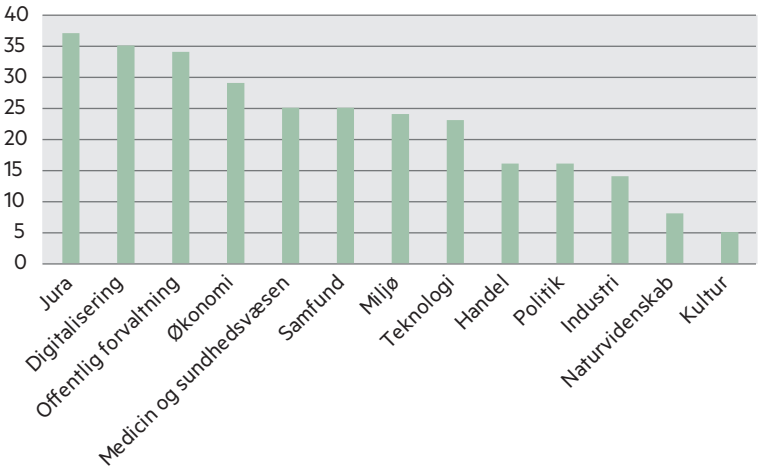
57 http://www.lr-coordination.eu/sites/default/files/Denmark/2018/ELRC%2B%20Workshop%20Report_Denmark.pdf

Deltagerne arbejdede overvejende med dansk, engelsk og tysk. Størsteparten arbejder med termsamlinger i en eller anden form. Interne termsamlinger ser ud til at være mest udbredt. Men der findes også en del virksomhedsinterne termbaser. Blandt termbankerne er det især EU's termbank IATE der bruges. Blandt de offentligt tilgængelige termbaser fremhæves den medicinske termbank SNOMED, socialebegreber.dk og sundhedsvæsenets begrebsbase. En interessant case der understreger behovet for samordning af terminologi på tværs af institutionerne, udgør universiteternes uddannelsesterminologi, hvor KU, AAU, CBS og AU har hver deres egen termbase. Her arbejdes der pt. på at skabe et fælles system hvilket vil kunne medføre en betydelig resursebesparelse.

Behovet for adgang til klare definitioner af begreber forekommer ikke blot blandt sprog- og kommunikationsmedarbejdere og sagsbehandlere, men i lige så høj grad blandt dataarkitekter, ingeniører, forretningsspecialister og it-udviklere. Oversættelse udgør fortsat det største område hvor der er behov for terminologi. Ca. 50 % af de adspurgte har angivet dette som den mest hyppige sammenhæng hvor der er behov for terminologi- eller begrebsarbejde. Men det er tydeligt at områder som drift og udvikling af it-systemer og forretningsudvikling vinder frem. De udgør sammenlagt 20 % af besvarelserne.

Både af besvarelserne af spørgeskemaet og af tilkendegivelserne i den efterfølgende bearbejdning af svarene på workshoppen fremgår det tydeligt at der er et stort behov for at få samlet terminologi og begreber og deres definitioner et fælles sted i en dansk termbank. De oplysningstyper der er størst behov for, er betydningsoplysninger (definitioner), ækvivalente termer (fx forkortelser, varianter og oversættelser til andre sprog) og oplysning om fagområde. Overraskende mange meldte også om behov for en strukturering af begreberne i forhold til hinanden i et begrebssystem, en taksonomi eller en tesaurus. Det er således ikke blot termerne og deres oversættelse der er behov for, men også information om deres indbyrdes sammenhæng.

Blandt de fagområder som hyppigst nævnes som dem der er størst behov for, er jura, digitalisering, offentlig forvaltning, økonomi og sundhed.



Workshoppens konklusioner

Der var bred enighed om at en dansk termbank vil

- styrke det danske fagsprog og det danske sprog generelt
- bidrage til bedre kommunikation både i den offentlige og i den private sektor
- medvirke til at sikre digitaliseringsklar lovgivning
- skabe bedre effektivitet i offentlige digitale systemer.

Workshoppens vigtigste anbefalinger var:

- Udviklingen af en dansk termbank skal være behovsdrevet, dvs. det skal være de fagområder og institutioner der har et konkret behov, der skal gå forrest med hensyn til at beskrive termer og lægge dem i termbanken

- Alle former for terminologi skal kunne finde plads i termbanken, så længe det er tydeligt markeret hvilken kvalitet og hvilken status termerne har
- Det offentlige kan med fordel være drivkraft for udviklingen, men både den offentlige og den private sektor bør få adgang til data og er velkomne til at bidrage med data
- Termbanken bør være åben for alle fra offentlige og private virksomheder til uddannelsesinstitutioner på alle niveauer og til den enkelte borger og have passende grænseflader til disse brugergrupper
- Termbanken skal administreres og kvalitetssikres af fagfolk, og der skal etableres et frugtbart samarbejde med de faglige ildsjæle rundt omkring i institutionerne
- Der skal skabes gode forbindelser mellem termbanken og den encyklopædiske viden der er beskrevet andre steder, fx i Den Store Danske Encyklopædi, i Trap Danmark, i Sprog- og Litteraturselskabets begrebsordbog og i semantiske beskrivelser som DANnet og FRAMEnet.


Folketinget.dk

8. Udvalgets anbefalinger

I dette afsnit præsenterer vi de anbefalinger som sprogteknologiudvalgets kortlægning af den aktuelle situation for dansk sprogteknologi har resulteret i.

Udvalget har i løbet af 2018 arrangeret i alt 6 workshops som fokuserede dels på forskellige interessentgrupper: brugere, udbydere og udviklere af sprogteknologi samt forskere og undervisere med sprogteknologi som speciale, dels på temaerne automatisk oversættelse og behovet for en national termbank. Hver workshop har haft mellem 15 og 30 deltagere, i alt ca. 120 repræsentanter for virksomheder, organisationer og offentlige institutioner, og er blevet suppleret med spørgeskemaer som især har haft til formål at afdække den aktuelle situation. Udvalgets anbefalinger er blevet præsenteret på et afsluttende seminar hvor alle workshopdeltagere har fået mulighed for at kommentere og supplere.

Som det er fremgået, er det tekniske og resursemæssige grundlag for udvikling af sprogteknologi for det danske sprog ikke på højde med sprogteknologien for hovedsprogene i de lande som vi ofte sammenligner os med (fx Norge, Sverige, Finland). På flere områder ligger dansk blandt de lavest rangerende sprog i Europa. Dette skyldes især fem forhold:

- Sprogsamfundets begrænsede størrelse og deraf følgende manglende attraktivitet som marked for sprogteknologi på dansk
- Sprogets særlige egenskaber (især den komplicerede fonetik og udtalestruktur)
- Mangel på visse basisresurser, især kvalitetsdata inden for tale, terminologi og semantik
- Manglende koordinering
- Fravær af en strategisk forskningsindsats og nedprioritering af uddannelser inden for dansk sprogteknologi i de seneste år.

På den anden side er det også fremgået at både den offentlige sektor og erhvervslivet udviser stor interesse i at der bliver udviklet god sprogteknologi for dansk. Det største potentiale ligger i at styrke samarbejdet mellem offentlige institutioner, små og store virksomheder og forskere bl.a. ved at gøre basisteknologier og resurser tilgængelige for alle og ved at sørge for at resurser og viden også fremover deles og opdateres.

Hvis regeringens målsætning om dansk sprogteknologi i verdensklasse skal virkeliggøres, er der derfor behov for:

- At koordinere indsatsen for dansk sprogteknologi, herunder tilvejebringe og udstille danske sprogresurser i en dansk sprogbank
- At understøtte danske virksomheders udvikling af sprogteknologi for dansk
- At gøre det attraktivt for udenlandske aktører at tilpasse deres produkter til dansk
- At styrke forskning og undervisning i dansk sprogteknologi.

Det er udvalgets vurdering at den mest effektive måde hvorved dansk sprogteknologi kan fremmes, er at indsatsen for dansk sprogteknologi koordineres, og danske sprogresurser i videst muligt omfang stilles frit til rådighed for udvikling og forskning på samme måde som der allerede i dag gives fri adgang til andre offentlige data, fx grunddata som oplysninger om personer, virksomheder, adresser, ejendomme, geografiske data mv.

Udvalgets kortlægning af området har vist at der i dag allerede eksisterer en grundstamme af resurser som vil kunne bringes i anvendelse hvis den nødvendige finansiering kan tilvejebringes. Der findes endvidere en lang række aktører som har den nødvendige ekspertise og er villige til at gøre en stor indsats for at udvikle og vedligeholde de resurser som mangler, for at udviklingen af dansk sprogteknologi kan drives frem til et højt kvalitetsniveau.

Initiativerne bør imidlertid suppleres med andre mere langsigtede initiativer inden for forskning og uddannelse som skal sikre at der også i fremtiden vil være de rette kompetencer til at drive udviklingen inden for dansk sprogteknologi videre.

Udvalget anbefaler:

1. Oprettelse af en organisation med ansvar for at etablere en dansk sprogbank og for at planlægge og igangsætte sprogteknologiske udviklingsprojekter

2. Sprogbanken skal tilvejebringe og vedligeholde danske sprogresurser som skal stilles til rådighed i høj lingvistisk kvalitet optimeret til sprogteknologiske formål.

   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.

Anbefaling 1 vedrører det organisatoriske niveau for sprogbanken, anbefalingerne i 2 omhandler de nødvendige resurser der skal tilvejebringes eller tilgængeliggøres i sprogbanken samt den teknologiske infrastruktur for sprogbanken. Anbefalingerne er meget konkrete og kan iværksættes straks under forudsætning af at den nødvendige finansiering er til stede. Anbefaling 3 og 4 omhandler de mere langsigtede kompetenceorienterede og forskningsmæssige forudsætninger for at der også i fremtiden kan udvikles sprogteknologi af høj kvalitet for dansk.

Initiativerne under anbefaling 1-2 kan løses og koordineres som en sammenhængende opgave og kan placeres under et enkelt ministerområde. Initiativerne under anbefaling 3 og 4 kræver samarbejde mellem flere ministerområder. Tre initiativer under anbefaling 1-2 kræver ligeledes samarbejde mellem ministerområderne. Disse tre initiativer er markeret særskilt som Anbefaling A-C.

Det bør tages i betragtning at udviklingen inden for kunstig intelligens og sprogteknologi går meget hurtigt, og at algoritmer og metoder hurtigt bliver forældede. Udvalget lægger derfor vægt på at de tiltag der foreslås, er langtidsholdbare og samtidig muliggør fleksibel tilpasning. Det handler derfor ikke om at tilvejebringe et stort antal forskellige resurser, men om at etablere lige præcis de resurser og datafødekæder som kan kickstarte udviklingen hen imod bedre sprogteknologi for dansk.

Der lægges således i høj grad op til at de foreslåede tiltag udvikles i tæt samråd med brugere og udviklere, og at indsatsen foregår kontinuerligt med løbende opfølgning og fokus på tiltagenes samfundsmæssige effekt.

8.1. Oprettelse af en organisation med ansvar for at etablere en dansk sprogbank

For at danske sprogresurser hurtigt og effektivt kan gøres tilgængelige og distribueres til virksomheder og offentlige institutioner, er der behov for at ansvaret for danske sprogresurser forankres et centralt sted.
Der findes allerede en lang række danske sprogresurser. En del af disse (især tekstkorpusser, taledata og annoterede datasæt) er utilgængelige for sprogteknologiske virksomheder på grund af ophavsretsmæssige begrænsninger og begrænsninger af hensyn til persondatasikkerheden. De få korpusser der er frit tilgængelige, er ofte uegnede som sprogteknologiske resurser da de ikke opfylder nutidens krav til volumen, opmærkning og aktualitet. Resurserne er typisk skabt i forbindelse med nationale og internationale offentlige forsknings- og udviklingsprojekter, men der har som regel ikke været bevillinger til at holde resurserne opdateret efter at projekterne er blevet afsluttet. Det betyder at de offentlige investeringer i udviklingerne af resurserne ikke har kunnet få deres maksimale effekt fordi resurserne ikke løbende er blevet vedligeholdt, videreudviklet og videreformidlet til udviklere og udbydere.

Andre danske sprogresurser er til en vis grad offentligt tilgængelige (især ordbøger og ordbaser samt terminologiske og semantiske resurser). Disse er typisk af høj lingvistisk kvalitet og kan umiddelbart videreudvikles som efterspurgt i markedet, evt. frikøbes og indgå i den samlede pakke af basisresurser. En tredje gruppe resurser forefindes hos private aktører som i et vist omfang vil kunne stille resurserne til rådighed på licensvilkår eller via frikøb.

Forskere, udviklere og virksomheder oplever det som særdeles vanskeligt og uoverskueligt dels at finde frem til de relevante date, dels at vurdere hvilken kvalitet og status de har.

Erfaringerne fra Finland viser at en decentral organisering af resurseindsamlingen kan føre til en uhensigtsmæssig spredning af ansvaret for resurserne og dermed mindre effektivitet og større omkostninger. En central placering er optimal både af hensyn til kontinuiteten og af hensyn til effektivt at kunne målrette en sådan tjeneste til erhvervslivet og den offentlige sektor.

- Udvalget anbefaler at det overordnede ansvar for danske sprogresurser placeres i en statslig styrelse, og at der i eller i tilknytning til denne styrelse oprettes en organisation der får til ansvar at etablere og drive sprogbanken
- Udvalget anbefaler at stat, regioner og kommuner informeres bedre om håndtering af sproglige resurser og om de muligheder der ligger i resurserne, både med henblik på at nyttiggøre dem til egne formål og med henblik på deling som frie resurser for at styrke udviklingen af bedre sprogteknologi for dansk.

Organisationens opgaver

Organisationen skal i samarbejde med relevante erhvervs- og forskningsmiljøer i Danmark og udlandet følge udviklingen inden for sprogteknologi og kunstig intelligens tæt og kunne tilpasse sine aktiviteter løbende. Organisationen skal være bemandet med tilstrækkelig administrativ, juridisk, sprogteknologisk og datalogisk kompetence.

Organisationens vigtigste opgaver er at

- etablere og drive sprogbanken
- at tilvejebringe de sprogteknologiske kvalitetsresurser som skal udstilles i sprogbankens resurseportal, herunder
  - at fastlægge formater og standarder for danske sprogresurser
  - at sørge for at allerede eksisterende resurser indsamles (evt. via frikøb eller licensaftaler) og udstilles et samlet sted
  - at udbyde, iværksætte og overvåge en række centrale resurseudviklingsprojekter
  - at sikre synergi og udveksling mellem udviklingsprojekterne
- at sørge for løbende vedligeholdelse af resurserne i takt med udviklingen i det danske sprogsamfund
- at sikre at de udviklede resurser deles effektivt
- at fastlægge en licensstruktur for resurser som ikke kan deles frit
- at bidrage til videndeling og kompetenceudvikling inden for sprogteknologi
- at sørge for rådgivning om sprogteknologi og vejledning i brugen af resurserne
- at etablere effektive kommunikations- og oplysningskanaler – herunder konferencer og kurser for brugerne
- at evaluere aktiviteternes effekt.

Organisationen kan med fordel løse opgaverne i partnerskab med andre organisationer og virksomheder og uddelegere delopgaver efter behov. Af hensyn til sikring af kvaliteten bør udvikling af nye resurser sendes i udbud i fri konkurrence, og de indkomne projektforslag evalueres af uafhængige eksperter.

Korpusser og datasæt skal udstilles i sprogbankens resurseportal, og det bliver en del af organisationens opgave at fastlægge standarder for metadata samt for annotation og sammensætning af tekstgenrer samt at

sørge for at etablere de juridiske, persondata- og it-sikkerhedsmæssige forudsætninger for driften af sprogbanken.

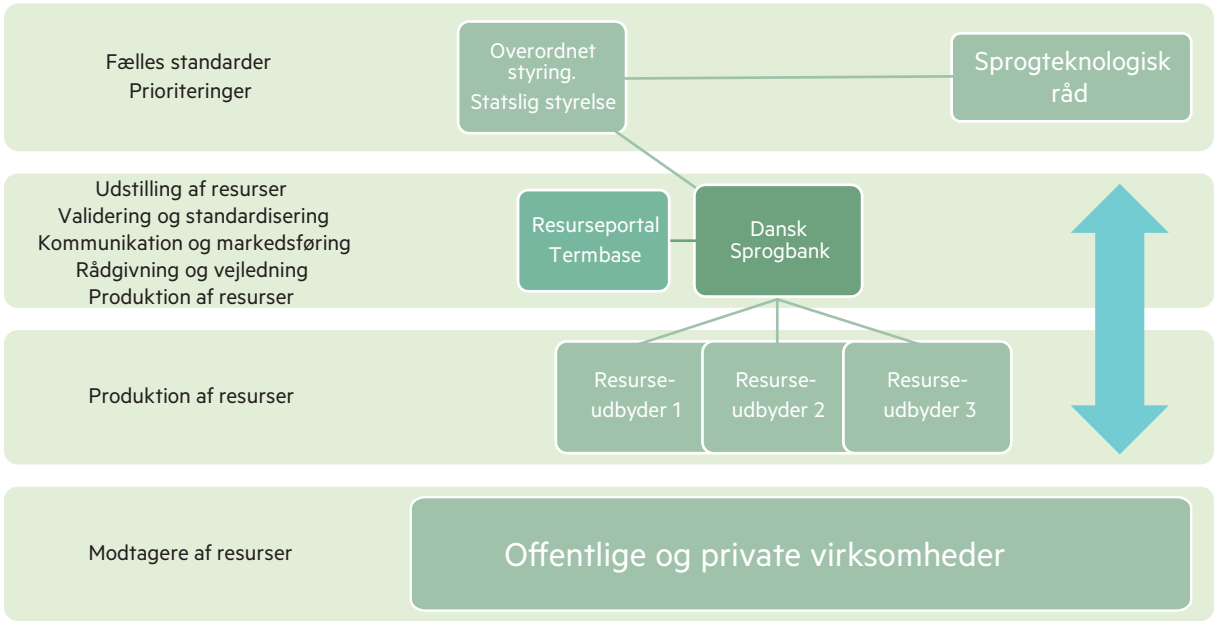
Et sprogteknologisk råd

Organisationen skal udføre sit arbejde i samråd med de centrale aktører som udvikler og forsker i dansk sprogteknologi, og med repræsentanter for brugere af sprogteknologi. Til dette formål oprettes et sprogteknologisk råd. Det sprogteknologiske råd får til opgave at bidrage til udvikling af en indsamlingsstrategi, planlægning af udviklingsprojekter og udvikling af resurseportalen og dermed sikre at sprogresurserne fremmer udviklingen og udbredelsen af dansk sprogteknologi mest muligt.

Udbud af sprogteknologiske resurseprojekter i åben konkurrence

Organisationen skal forvalte de midler som afsættes til sprogteknologi og udbyde de aftalte sprogteknologiske resurseprojekter og demonstrationsprojekter i åben konkurrence og med uvildig bedømmelse af internationale fageksperter i lighed med de udbudsprocesser der gælder for de danske forskningsråd og tilsvarende initiativer i Nederlandene og Island.

Forslag til organisering af sprogteknologiindsatsen



8.2. En dansk sprogbank

Anbefalingen består af 6 initiativer:

1. Et tidskodet dansk talesprogskorpus
2. En sprogteknologisk værktøjskasse
3. Danske tekstkorpusser og opmærkede guldstandarder
4. En avanceret dansk ordbase
5. En dansk termbank
6. En resurseportal til distribution og deling af sprogresurser.

For at talegrænseflader til digitale assistenter og robotter skal fungere optimalt på dansk, er der først og fremmest behov for en taleresurse af høj lingvistisk kvalitet som gør det muligt at skabe en robust talegenkendelse og varieret syntetisk tale. Som udgangspunkt er det tilstrækkeligt at udvikle en generel, grundigt annoteret og kvalitetssikret taleresurse der kan danne basis for udvikling af mere specifikke resurser.

Udvalgets undersøgelser har vist at kvaliteten af dansk talegenkendelse og talesyntese ikke når et tilfreds-

stillende niveau fordi såvel de store kommercielle udenlandske aktører som de små danske taleteknologivirksomheder oplever at initialomkostningerne til tilvejebringelse af en sådan resurse er for store. Der vil være betydelige omkostninger sparet ved at basisresursen er frit tilgængelig i stedet for at hver virksomhed skal producere den fra grunden. Basisresursen vil give virksomheder muligheden for at foretage et teknologispring gennem tilpasning og videreudvikling af basisresursen til nye specialiserede kommercielle og ikke-kommercielle produkter og tjenester.

Udvalgets undersøgelser har også vist at der i stort omfang savnes bedre værktøjer og resurser til at behandle de oplysninger der ligger i tekster lige fra stavekontrol over automatisk resumering, identifikation af navne og steder, klassifikation af dokumenter og emneområder eller analyse af de holdninger der kommer til udtryk i fx kommentarer på Facebook eller i en chatbot. De resurser og værktøjer der skal bruges til at få det maksimale udbytte af tekster, er de samme der skal bruges for at et system kan gå videre fra den umiddelbare forståelse af tale til en egentlig forståelse af hvad der bliver sagt, og til omsætning af det sagte til handlinger eller generering af svar i en dialog. Der er med andre ord brug for resurser der kan fremme systemernes sprogforståelse, dvs. tekstsamlinger, annoterede datasæt m.m.

De gængse teknikker inden for kunstig intelligens med træning af algoritmer på store tekstmængder kan give gode resultater en del af vejen og til visse typer af applikationer, men rå tekst er ofte ikke tilstrækkelig når der skal udvikles mere avancerede applikationer, fx systemer der kan indgå i dialog med en bruger eller foretage et selvstændigt ræsonnement. Her er der brug for tekstsamlinger/datasæt som er annoteret med lingvistisk information, og for adgang til information om ordenes betydning (semantik) og kombinationspotentiale.

En særlig resurse udgør i den forbindelse tekster som foreligger på flere sprog, typisk oversatte tekster, idet de er særligt velegnet til træning af automatiske oversættelsesværktøjer. Her er det især indsamlingen af teksterne der udgør en udfordring simpelthen fordi der hos brugerne ikke er tilstrækkelig opmærksomhed på at oversatte tekster kan bruges til at træne et automatisk oversættelsessystem.

Også en national termbank vil kunne fungere som en sprogresurse først og fremmest fordi den vil kunne indeholde de specialiserede fagudtryk og -begreber som i stadigt større omfang allerede indsamles og beskrives for at sikre kvalitet, sikkerhed og funktionalitet i forbindelse med offentlige it-projekter, fx i Socialstyrelsen og Sundhedsstyrelsen.

Der er en række andre datasæt der på lidt længere sigt bør indgå i resurseportalen. Store nationale kulturarvsprojekter som fx Gyldendals Encyklopædi og Trap Danmark er særdeles interessante og relevante resurser af høj kvalitet som bør være tilgængelige som datasæt med henblik på udvikling af sprogteknologi. Det kongelige Bibliotek har igennem tiden digitaliseret store mængder avistekster, bøger og andre tekstresurser. Også DR og Danske Medier er i besiddelse af store datamængder som bør kunne bringes i spil og styrke udviklingen af sprogteknologi på dansk.

8.2.1. Et tidskodet dansk talesprogskorpus

For denne resurse kan der gives mere konkrete rammer da ethvert projekt om talegenkendelse (den mest datakrævende af de almindelige taleteknologier) skal bruge materialer af omtrent samme tilsnit.

For at der kan udvikles robust og fleksibel talegenkendelse der kan behandle tale fra en vilkårlig dansker, skal der trænes på et materiale af minimum 200 timers taleoptagelser (200-2000 talere) med systematisk variation af køn, alder og herkomst. Hver taler skal bidrage med både spontan og planlagt tale, samt varierende talestil (hastig vs. tydelig). Hver taleoptagelse skal dække alle det danske sprogs fonetiske elementer.

Fremstillingen af et sådant talesprogskorpus har en række faser hvoraf stemmeoptagelserne kun er den første. Den klart mest omkostningstunge del følger efter optagelserne idet talematerialerne - for at have værdi som sprogteknologisk komponent - skal annoteres med lydskrift og tidskodes af fonetiske eksperter. Denne aktivitet tager typisk 20 gange så lang tid som lyden selv repræsenterer. Et talekorpus på 200 timer vil således tage 4000 timer at efterbehandle.

Lydoptagelserne skal transskriberes både ortografisk og fonetisk. Dette kan ikke gøres automatisk, da uundgåelige diskrepanser mellem tekst og oplæsning (for indlæst tale) og uregistrerede fonetiske reduktioner ville forurene korpusset og ødelægge dets værdi for maskinlæringen.

Det anbefalede talesprogskorpus vil gøre det muligt at træne sprogteknologiske løsninger til intelligent taleforståelse, som for eksempel talegenkendelse, talesyntese, taleridentifikation, IVR (interactive voice-response) og dialogsystemer.

8.2.2. Danske tekstkorpusser og opmærkede guldstandarder

Udvalgets undersøgelser har vist at dansk sprogteknologi i høj grad mangler frit tilgængelige tekstkorpusser der har størrelse og kvalitet så de kan bruges fx til emneklassifikation, sentimentanalyse, automatisk tekstresumering, indholdsbaseret søgning, maskinoversættelse, dialogmodellering m.m.

Løbende indsamling af tekstdata

Omfattende tekstkorpusser er essentielle grundressourcer i sprogteknologisk sammenhæng da megen moderne sprogteknologi er baseret på maskinlæringsteknologier, herunder deep learning, der for at fungere godt skal trænes på enorme mængder af tekstmateriale fra et relevant domæne. Korpusser skal indsamles, renses og opmærkes løbende. Ellers vil de hurtigt blive uaktuelle da der konstant kommer mange nye ord og vendinger til i takt med samfundsudviklingen.

- Udvalget anbefaler at der med udgangspunkt i allerede eksisterende korpusresurser etableres en kontinuerlig indsamling, rensning og opmærkning af tekstkorpusser fra forskellige almensproglige og fagsproglige domæner.

Kvalitetsopmærkede guldstandarder

Opmærkning af tekster med lingvistiske informationer af høj kvalitet er nødvendige i maskinlæringsteknologier for at opnå den optimale kvalitet. Der er derfor behov for en række kvalitetskorpusser med kontrolleret opmærkning – såkaldte guldstandarder – som kan danne et solidt grundlag for træning af sprogteknologiske komponenter i danske sprogteknologiske virksomheder.

- Udvalget anbefaler at der efter 3 år skal foreligge et antal frit tilgængelige korpusser der er egnet som guldstandarder i professionelle maskinlæringsprojekter, fx modellering af dialoger, syntaksanalyse (træbanker) og avanceret betydningsanalyse, fx holdningsanalyser (sentiment-analysis og opinion-mining).

Frigivelse af flere tekster til sprogteknologiske formål

Den praktiske udnyttelse af de tekstresurser der findes for dansk, er ofte umuliggjort af hensyn til copyright (fx forlagspublikationer, avisartikler og TV-tekstning), klientbeskyttelse (fx sagsbehandlingsakter fra kommuner, regioner og politiet) og privatlivsbeskyttelse (fx sociale medier, blogs og e-mailkorrespondance).

- Udvalget anbefaler en ændring af reglerne for pligtaflevering til Det Kgl. Bibliotek sådan at tekstindholdet i afleverede materialer kan anvendes frit til udvikling af sprogteknologi, dog således at ophavsretten for andre anvendelser bevares fuldt ud (Anbefaling A).

Det skal i den forbindelse naturligvis sikres at tekstejerens ophavsrettigheder mv. ikke forringes. Dette kan fx opnås ved at teksterne udstilles i en fragmenteret form sådan at længere afsnit og følsomme oplysninger ikke kan rekonstrueres. En sådan maskeret tekstversion vil ofte være fuldt tilstrækkelig til de fleste maskinlæringsformål.

Indsamling af tekstmateriale som er frit tilgængeligt eller tilvejebringes i offentligt regi eller i offentligt finansierede projekter

Også andre metoder til indsamling af frit tilgængelige tekstresurser skal undersøges. Mange danske tekster er allerede uden copyright og kan uden videre inkluderes (litteratur af ældre dato, lovsamlinger og forordninger, referater fra byråd og Folketinget, domme, undervisningsmaterialer til skole og uddannelser, m.m.), lige som der findes fora på internettet hvor danske tekster publiceres med eksplicit afståelse af rettigheder. I mange tilfælde har kommuner og virksomheder udlånt store tekstmaterialer til brug for forskning og udvikling, og de vil måske kunne tillade fri benyttelse af maskerede versioner.

- Udvalget anbefaler at indsamlingen af tekster som produceres i offentlig regi eller i offentligt finansierede projekter, sættes i system. (Anbefaling B)

Store nationale kulturarvsprojekter som fx Gyldendals Encyklopædi og Trap Danmark er ligeledes interessante og relevante resurser af høj kvalitet som bør være tilgængelige som datasæt med henblik på udvikling af sprogteknologi. Også DR og Danske Medier er i besiddelse af store datamængder som bør kunne bringes i spil og styrke udviklingen af sprogteknologi på dansk.

- Udvalget anbefaler at store nationale kulturarvsprojekter stiller deres data til rådighed for sprogteknologi.

Flersprogede tekster og automatisk oversættelse

Korpusindsamlingen bør også omfatte danske tekster oversat til fremmede sprog samt udenlandsk tekst oversat til dansk (organiseret som såkaldte parallelkorpusser) til træning af maskinoversættelsessystemer. Specielt i forhold til indsamling af flersprogede tekster med henblik på automatisk oversættelse konstaterer udvalget at der i den offentlige sektor er behov for en større bevidsthed om hvilken betydning oversatte tekster har for udvikling af automatisk oversættelse og andre sprogteknologiske applikationer. Der er behov for en fælles klassifikationsmodel/metadata for oversatte tekster og oversættelseshukommelser som gør det muligt at vurdere om teksterne egner sig til deling, dvs. ikke indeholder persondata eller andre klassificerede oplysninger, og hvilke fagområder og teksttyper de dækker. Der er behov for klare regler for hvem der har dispositionsretten over de oversatte data i forbindelse med udbud af oversættelsesydelser, fx er det i dag ikke en forpligtelse for private udbydere at overdrage oversættelseshukommelsen til den offentlige institution eller til en resurseportal.

Der er endvidere behov for gode analyser af hvordan det oversatte materiale i øvrigt kan bringes i spil, fx i forbindelse med udvikling af automatisk oversættelse af websider m.m., som resurse for tolkning eller ekstraktion af flersproget såvel som dansk terminologi, ekstraktion af flerordsudtryk og faste vendinger mv.

- Udvalget anbefaler at eksisterende oversatte tekster eller oversættelseshukommelser identificeres, klassificeres og deles som en del af den sprogteknologiske resurseportal
- Udvalget anbefaler at der foretages en generel kortlægning af brugen af oversættelsesopgaver i offentligt regi, specielt med henblik på volumen, udbudsregler, arbejdsgange og arkivering, og at der udvikles en model for offentligt-privat samarbejde hvor både offentlige og private aktører samarbejder om løbende at bidrage til indsamling af oversættelsesresurser (Anbefaling C).

Korpusser for dansk tegnsprog

Dansk tegnsprog er et selvstændigt sprog i Danmark. Ifølge Danske Døves Landsforbund (DDL) findes der ca. 4.000 døve personer i Danmark som har dansk tegnsprog som modersmål (tallet er fra 2014). Det vil sige at lidt under en promille af den danske befolkning har dansk tegnsprog som modersmål. Dertil kommer et antal hørende familiemedlemmer og professionelle der bruger dansk tegnsprog i deres kontakt med døve.

For at kunne understøtte udvikling af kommunikationsmidler til mennesker der bruger dansk tegnsprog, er der brug for at dansk tegnsprog løbende dokumenteres i form af transskriberede og annoterede videooptagelser. Der findes allerede i dag metoder til automatisk gengivelse og genkendelse af tegnsprog, men de kræver et velstruktureret og annoteret korpus for at blive anvendt på dansk tegnsprog.

- Udvalget anbefaler at der sideløbende med indsamling af tekster og datasæt for dansk også iværksættes initiativer der understøtter opbygningen af et korpus for dansk tegnsprog.

8.2.3. En avanceret dansk orddatabase

En avanceret dansk orddatabase vil kunne danne fundament for alle de øvrige udviklede sprogresurser, fx også til annotering af tekster.

Alle leksikalske indgange (lemmaer) skal annoteres for fx:

- Staveformer (normerede og faktisk forekommende)
- Ordklasse og bøjning
- Udtalevariation (både leksikalsk-fonetisk form og realisering i flydende tale)

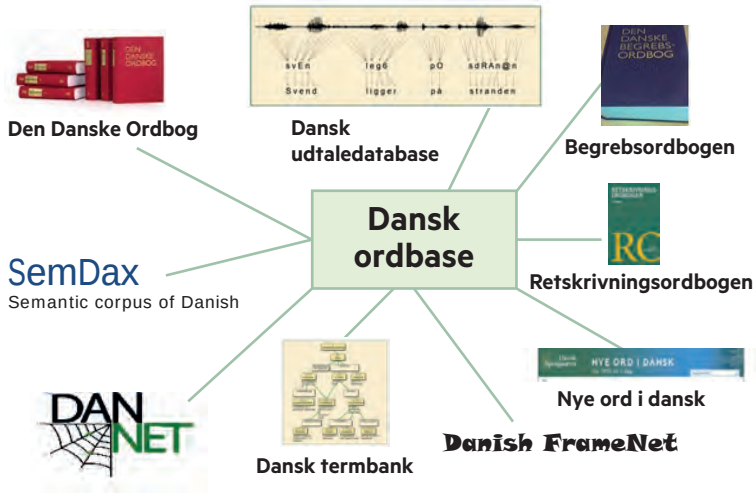
- Sammensætningspotentiale
- Emneklassifikation (overordnet domæne)
- Anvendelser i faste vendinger
- Syntaksoplysninger i form af valens- og dependensinformation
- Semantiske oplysninger fx i form af betydninger, positiv-negativ konnotation, semantiske typer, roller og relationer.

Ordbasen skal tage udgangspunkt i de ordbøger over det danske sprog som allerede findes i en standardiseret og struktureret form især Sprogteknologisk Ordbog, Det danske wordnet (DANnet), det danske FrameNet, Retskrivningsordbogen og Nye ord i dansk fra Dansk Sprognævn og Den Danske Ordbog og Begrebsordbogen fra Det Danske Sprog- og Litteraturselskab. Der skal indgås passende aftaler med ordbogsleverandørerne om brugen af ordbøgerne og om løbende supplering med nye ord. De leksikalske indgange skal endvidere kobles til en dansk udtaledatabase.

Ordbasen skal baseres på åbne formater og internationale standarder og være designet til at modtage nye data udtrukket fra løbende indsamlet, nyt tekstmateriale. Ordbasen skal endvidere omfatte regler, templates og algoritmer til prædiktion af ukendte ords ordklasse, udtalemuligheder, betydning, sammensætningspotentiale, m.m.

Ordbasen skal sætte standarden for de leksikalske beskrivelser i de øvrige sprogresurser og dermed sikre at der kan komme sammenhæng i de sprogteknologiske løsninger.

- Udvalget anbefaler at eksisterende leksikografiske resurser samordnes i en ordbase som en rigt struktureret leksikalsk database der dækker det (nutidige) almene danske ordforråd samt det ordforråd man typisk finder i danske fagtekster inden for en række professioner (fx jura, medicin og økonomi). Desuden skal egennavne (proprier) dækkes, herunder danske stednavne og personnavne.
- Udvalget anbefaler at der etableres links mellem ordbasen og den danske termbank således at man ved opslag af et almensprogligt ord i ordbasen kan få adgang til de oplysninger om ordet der er registreret som term i termbanken, således at man også kan få oplysninger om mere specifikke termer.



Eksisterende og fremtidige resurser som vil kunne indgå i en avanceret dansk orddatabase. Kun udtalebasen og termbanken er nye resurser. Resten er eksisterende resurser der skal integreres og videreudvikles løbende.

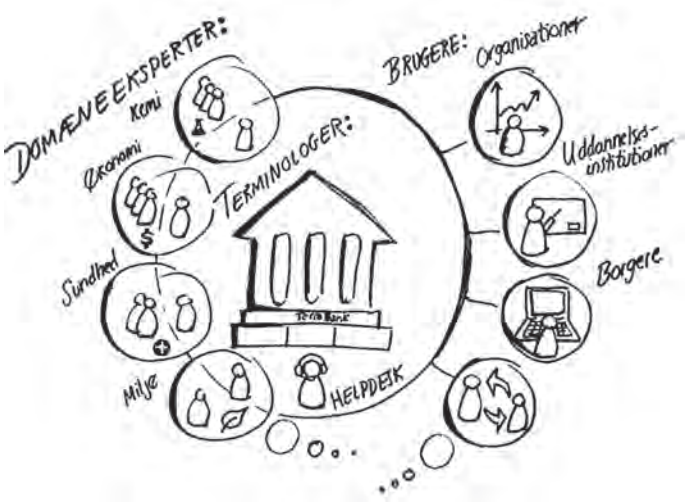
8.2.4. En dansk termbank

- Udvalget anbefaler at der oprettes en dansk termbank der giver adgang til det danske fagordforråd.
  - Termbanken skal først og fremmest gøre de begreber og definitioner tilgængelige der modelleres i offentligt regi og i forbindelse med arbejdet med den fællesoffentlige digitale arkitektur, og udstille dem i en lettilgængelig form i en online-database

- Termbanken skal endvidere samle oplysninger om andre begrebssystemer, klassifikationer, terminologilister mv. så de kan tilgås fra et fælles adgangspunkt af fx undervisningsinstitutioner, tolke i sundheds- og retsvæsenet samt danske virksomheder
- Udviklingen af termbanken skal trække på de erfaringer der allerede er indhøstet med begrebsarbejdet i fx Socialstyrelsen og Sundhedsstyrelsen, inddrage de ekspertfora og netværk der allerede er dannet inden for den offentlige sektor, og yderligere understøtte udvekslingen af erfaringer og ideer
- Termbanken skal endvidere opbygge ekspertise i rådgivning af offentlige institutioner der skal i gang med modellering af forretningsområder.

Termbanken vil gøre det muligt at træne en lang række sprogteknologiske løsninger til automatisk emneklassifikation og sikring af terminologisk konsekvens i tekstproduktion og kommunikation. Adgangen til begrebsmodeller for et bestemt fagområde kan få afgørende betydning for udvikling af dansk sprogteknologi og understøtte udvikling af mange former for applikationer inden for de forskellige fagområder.

Terminologiske data kan give adgang til det centrale faglige ordforråd med systematisk betydningsangivelse, til relationer mellem begreber til udvikling af sprogforståelse, til automatisk tekstklassifikation på basis af domænemodeller og til understøttelse af automatiske ræsonnementer og intelligent dialog.



Udviklingen af termbanken kan med fordel tage udgangspunkt i eksisterende forarbejder, fx resultaterne af projektet DANTERMBANK som med en bevilling på 5 mio. kr. fra Velux Fonden udviklede et koncept for en dansk termbank.

8.2.5. Indsamling og/eller udvikling af sprogteknologiske værktøjer

Sprogteknologiske værktøjer bearbejder typisk tekster i flere trin og med flere formål. En helt grundlæggende funktion er fx automatisk at opdele teksten i relevante enheder, fx ord eller sætninger. Allerede her kan man støde på vanskeligheder, fx ved analyse af datoer, adresser eller forkortelser hvor punktummer ikke signalerer en sætningsgrænse. Endvidere skal man kunne afgøre om flere ord udgør en helhed fx "social sikring" eller "Det centrale Personregister". Der eksisterer allerede i et vist omfang sprogteknologiske værktøjer der kan løse disse problemer, men de skal videreudvikles/opdateres og tilgængeliggøres. Visse værktøjer er endvidere belagt med restriktioner der gør det vanskeligt for virksomheder at bringe dem i anvendelse.

- Udvalget anbefaler at der tilvejebringes og udvikles sprogteknologiske værktøjer der kan understøtte behandlingen af dansk tekst og tale.

8.2.6. En resurseportal til distribution og deling af sprogresurser

De sprogteknologiske sprogresurser skal distribueres effektivt til alle interesserede. Det kræver en velfungerende infrastruktur, en sprogteknologisk resurseportal. Det skal undersøges i hvilket omfang eksisterende dataportaler såsom DK-CLARIN kan indgå i eller danne udgangspunkt for portalen, og hvordan data fra andre

danske og internationale portaler, fx Open-Data-dk, datafordeler.dk, META.share, ELRC, ELRA-ELDA m.fl., kan integreres. Endvidere bør EU's planer om oprettelse af fælleseuropæiske dataportaler inddrages.

- Udvalget anbefaler at der etableres en resurseportal der udstiller resurserne i de formater som skønnes bedst egnet fx i et repositorium, via API'er, i databaser mv. et centralt sted
- Udvalget anbefaler at der udvikles en langsigtet strategi for løbende indsamling og udstilling af tekst-, term- og talematerialer i den takt de modtages, fx i samarbejde med eksterne donorer.

Portalen skal endvidere indeholde et arkiv til manualer og resursebeskrivelser, et bibliotek af relevant litteratur, et brugerkartotek med kontaktoplysninger og et forum til udveksling af nyheder, spørgsmål og svar.

8.3. Styrkelse af kompetenceudvikling og uddannelse inden for dansk sprogteknologi

Siden 2010 er en række datalingvistiske og sprogteknologiske uddannelser blevet nedlagt. Der findes i dag ingen samlet uddannelse i dansk sprogteknologi på danske universiteter. Hertil kommer at de uddannelser hvori sprogteknologi indgår, som udgangspunkt ikke beskæftiger sig med danske sprogdata, men udelukkende med engelske data. Det har flere årsager: Dels er flere engelske data og værktøjer gratis tilgængelige for underviserne, dels undervises der på engelsk, dels har mange af de studerende ikke dansk baggrund.

Der opbygges derfor i dag ikke i tilstrækkeligt omfang viden om dansk sprogteknologi og brugen af danske data.

- Udvalget anbefaler at der nedsættes en arbejdsgruppe bestående af den sprogteknologiske organisation, Uddannelses- og Forskningsministeriet og uddannelsesinstitutionerne der får til opgave at undersøge mulighederne for at:
  - udvikle tilbud om undervisning i dansk sprogteknologi med særlig vægt på sprog og danske data på mindst to universiteter hvor der undervises i data science og kunstig intelligens, fx som et supplement til uddannelsen i it og kognition på KU, i data science på ITU eller i kunstig intelligens og big data på DTU eller på nogle af IT-uddannelserne på AU og AAU
  - udvikle efteruddannelsestilbud i håndtering af danske data og i natural language processing som enkeltfag eller korte kurser for dataloger
  - skabe en bedre forbindelse mellem de datalogiske, sprogteknologiske og lingvistiske undervisningsmiljøer på universiteterne for at udvikle relevante uddannelsestilbud
  - oprette en bachelor og en kandidatuddannelse i dansk med specialisering i sprogteknologi
  - øremærke midler til ph.d.-stipendier inden for feltet, fx inden for taleteknologi, tekstforståelse og semantik, udvikling og opmærkning af resurser, dialogmodellering og vidensmodellering
- Udvalget anbefaler at sprogteknologi indgår som et element i it-undervisningen i grundskolen og ungdomsuddannelserne. Sprogteknologi er fx velegnet som tema for tværfaglig undervisning (fx sprog og it).

8.4. Styrkelse af forskning i dansk sprogteknologi

En forudsætning for at dansk sprogteknologi når et niveau der nærmer sig verdensklasse, er at der sker et markant løft af forskningen inden for dansk sprogteknologi. Der har igennem de sidste 20 år været afsat relativt få forskningsmidler til dansk sprogteknologi, fx til udvikling af en forskningsinfrastruktur (DK-CLARIN), til en sprogteknologisk ordbog, til udvikling af talegenkendelse og til udvikling af semantiske netværk (DanNet, FrameNet). Midlerne er kommet fra forskellige kilder i forskningssystemet og fra fonde, og resursetildelingen har ikke været koordineret eller strategisk forankret.

Etableringen af en dansk sprogbank og en koordineret indsats for at tilvejebringe de nødvendige basisresurser udgør kun et første skridt. Hvis dansk sprogteknologi skal kunne måle sig med de førende i verden, skal der ydes en målrettet forskningsindsats på en række områder, fx

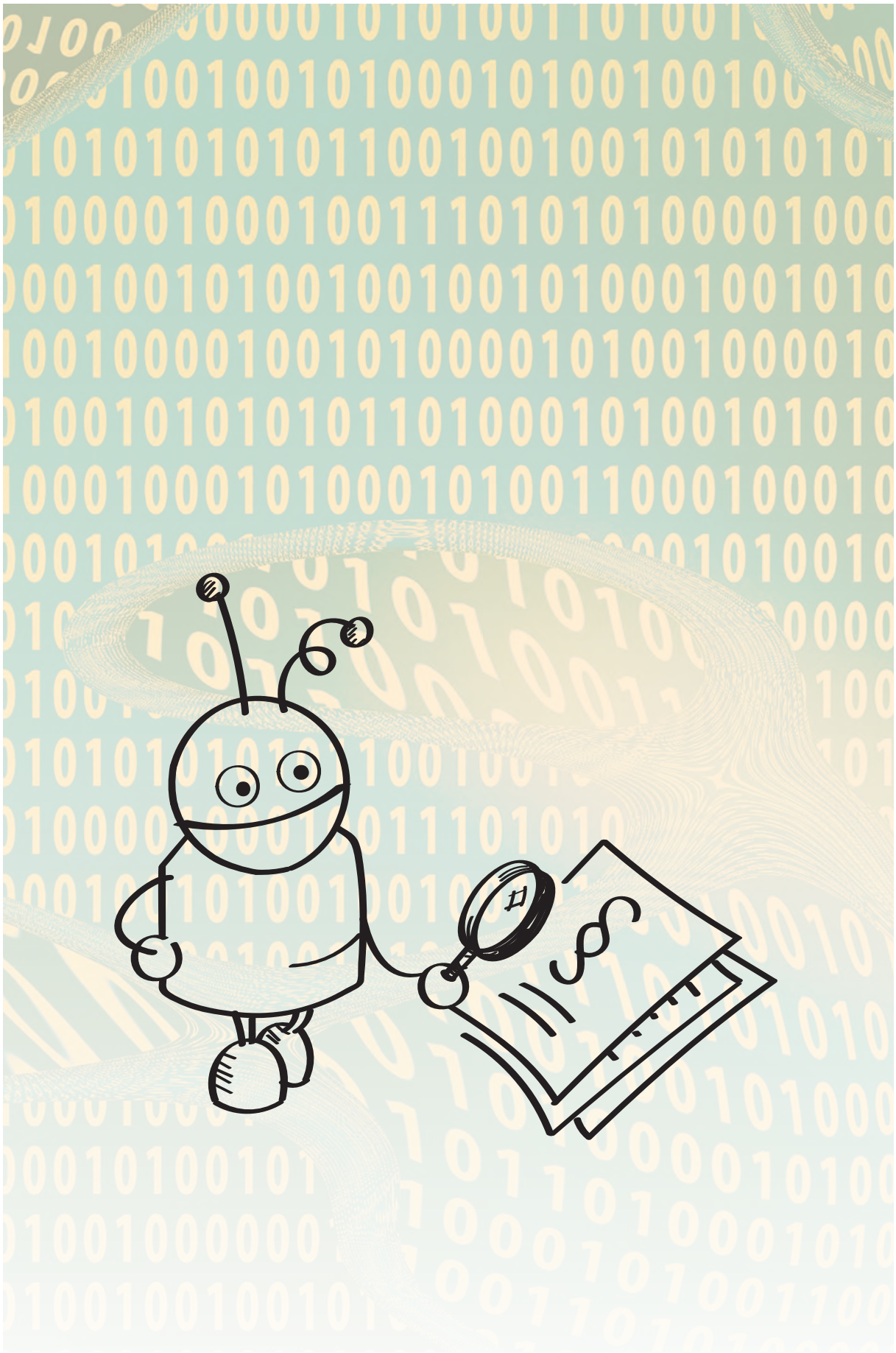
- Forskning i tryk, prosodi mv. på dansk for at kunstige stemmer lyder naturligt
- Forskning i stemmeføringen i dialog og i dialogstruktur og -grammatik

- Forskning i robust sprogforståelse i forbindelse med ufuldstændige ytringer, forkerte formuleringer, fejlstavninger, afbrydelser, rettelser mv
- Forskning i hvordan de sproglige resurser kan anvendes helt konkret i brugerrettede applikationer, og hvordan brugerne reagerer og interagerer sprogligt med systemerne
- Forskning i hvordan man mest effektivt opbygger taleteknologi og sprogforståelse i nye domæner
- Forskning i hvordan man effektivt kan udtrække terminologi og anden sproglig information
- Forskning i hvordan man automatisk eller semiautomatisk kan opbygge ordnet og begrebsmodeller
- Forskning i hvordan man kan udvikle bedre sprogteknologi til mennesker med behov for kommunikationshjælpemidler
- Forskning i hvordan man kan udvikle sprogteknologi for dansk tegnsprog, fx en automatisk oversætter fra dansk tegnsprog til dansk.

- Udvalget anbefaler at der afsættes selvstændige forskningsresurser til dansk sprogteknologi, og at de bør opslås i fri konkurrence efter en samlet plan der afstemmes med relevante aktører

- Udvalget anbefaler at der etableres tenure track-ordningen for at tiltrække forskere med interesse for dansk sprogteknologi.

Relevante aktører kunne fx være den sprogteknologiske organisation, Den Frie Forskningsfond, Grundforskningsfonden, Innovationsfonden, Danske Universiteter m. fl. som sætter fokus på at der i fri konkurrence tilvejebringes nye forskningsresultater der kan sikre anvendelsen af dansk i sprogteknologiske applikationer og applikationer der bygger på kunstig intelligens og sprog.



9. Forslag til finansiering

9.1. Udgifter fordelt på anbefalinger

I det følgende præsenteres et overslag over de udgifter som udvalgets anbefalinger vil medføre.

Udgifterne under 1 og 2 udgør det samlede finansieringsbehov for sprogbanken, i alt 55,6 mio. kr. fordelt over 4 år fra 2020 til 2023. Udgifterne under 3 og 4 forudsætter at initiativerne helt eller delvist finansieres af andre ministerområder, især Uddannelses- og Forskningsministeriet, samt forskningsfonde og forskningsråd.

Økonomiske forudsætninger for en sprogteknologisk basisorganisation					
	Etablering			**Overgang til drift**	
i 1.000 kr.	**2020**	**2021**	**2022**	**2023**	**i alt**
1. Oprettelse af en organisation med ansvar for at etablere en dansk sprogbank og for at planlægge og igangsætte sprogteknologiske udviklingsprojekter	4.000	4.000	4.000	3.000	15.000
2. Udvikling af en række danske sprogresurser i høj lingvistisk kvalitet optimeret til sprogteknologiske formål som skal stilles til rådighed i sprogbanken	11.250	12.750	9.800	6.800	40.600
2.1. Et tidskodet dansk talesprogskorpus	1.500	1.500	300	300	3.600
2.2. En sprogteknologisk værktøjskasse	1.000	500	500	500	2.500
2.3. Danske tekstkorpusser og opmærkede guldstandarder samt et korpus for dansk tegnsprog	3.750	5.250	4.000	1.000	14.000
2.4. En avanceret dansk ordbase	1.000	2.500	2.000	2.000	7.500
2.5. En dansk termbank	2.000	2.000	2.000	2.000	8.000
2.6. Etablering og drift af en resurseportal til distribution og deling af sprogresurser i sprogbanken	2.000	1.000	1.000	1.000	5.000
1-2 Sprogbanken i alt	**15.250**	**16.750**	**13.800**	**9.800**	**55.600**
3. Styrkelse af kompetenceudvikling og uddannelser inden for dansk sprogteknologi	500	2.000	2.000	2.000	6.500
4. Styrkelse af forskning i dansk sprogteknologi	500	5.000	10.000	10.000	25.500
1-4 Alle initiativer i alt	**16.250**	**23.750**	**25.800**	**21.800**	**87.600**

I forslaget er der taget højde for at visse ordbogsresurser allerede er til stede og løbende opdateres, fx Retskrivningsordbogen, og at der er opnået finansiering til fx digitalisering af Den Danske Begrebsordbog og udvidelse af DanNet via fondsmidler, samt at en række korpusser allerede foreligger fx i DK-CLARIN og andre portaler. Der er afsat beløb til harmonisering, standardisering og sammenkædning af eksisterende resurser og især til opmærkning af guldstandarder som er det mest omkostningstunge. De eksisterende tekstresurser som foreslås prioriteret, kunne være Retsinformation, borger.dk, Folketingstidende, allerede eksisterende fagkorpusser og almensproglige korpusser. Blandt korpusser som skal indsamles fra bunden, kan nævnes et dialogkorpus, et talesprogskorpus og fagkorpusser for nye domæner.

Udvalget har ikke set det som sin opgave i detaljer at pege på hvilke resurser der skal prioriteres, men har taget de eksisterende resurser og de resurser som er blevet foreslået i forbindelse med udvalgets workshops, som udgangspunkt for de økonomiske beregninger. Det vil være en opgave for sprogbanken og det sprogteknologiske råd at foretage de endelige prioriteringer.

Udvalget har endvidere ikke taget stilling til den løbende drift efter etableringsperioden, men har i beregningerne forudsat at nogle af resurserne vil være overgået til driftsfasen i det sidste af overslagsårene.

Supplerende anbefalinger	
A. Ændring af regler for pligtaflevering	Initiativet foreslås placeret i og finansieret af Kulturministeriet
B. Udvikling af retningslinjer for indsamling af tekster produceret i offentlige institutioner eller i offentligt finansierede projekter	Initiativet forelså placeret i og finansieret af Moderniseringsstyrelsen eller Innovationsministeriet
C. Kortlægning af oversættelsesopgaver i det offentlige og revision af udbudsreglerne mht. rettigheder til data og udvikling af en ny model for offentligt-privat samarbejde om oversættelse	Initiativet foreslås placeret i og finansieret af Moderniseringstyrelsen

Sprogbanken vil kunne deltage i arbejdet med at virkeliggøre de supplerende initiativer, men det overordnede ansvar bør placeres i de ovenfor nævnte ministerier eller styrelser. Det vurderes at aktiviteterne vil kunne rummes i de respektiver ministeriers budgetter.

9.2. Det samfundsmæssige potentiale

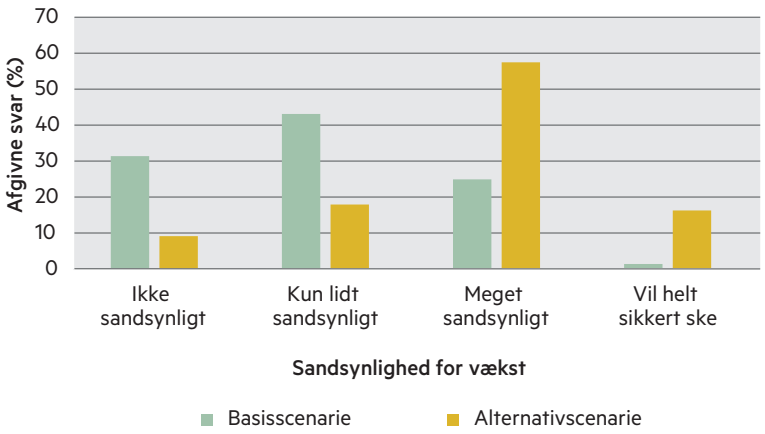
Hvilken samfundsøkonomisk effekt vil de anbefalede tiltag kunne opnå på dansk sprogteknologi i de nærmeste år fremover? Vi fremlægger her et estimat baseret på indikatorer uddraget af den samlede feedback fra workshops og spørgeskemaundersøgelser for alle interessegrupper.

Vores analyse tager udgangspunkt i to scenarier, basisscenariet (hvor ingen af udvalgets anbefalinger følges) og alternativscenariet (hvor alle anbefalinger følges). Først undersøges det i hvilken grad interessegrupperne forventer en forskel for udviklingen af dansk sprogteknologi i basisscenariet og i alternativscenariet, eller med andre ord, om de anbefalede tiltag vil virke stimulerende eller ej. Derefter vurderer vi det økonomiske volumen af det nuværende sprogteknologiske marked i Danmark. Baseret på disse to delresultater fremskriver vi den samfundsøkonomiske udvikling i henholdsvis basisscenariet og alternativscenariet. Forskellen mellem disse to udviklinger udgør vores samlede estimat af de anbefalede tiltags samfundsøkonomiske effekter.

Stagnation eller vækst

Vil markedet for dansk sprogteknologi vokse eller stagnere i de næste år? For at belyse branchens egne forventninger til væksten i basisscenariet og i alternativscenariet blev alle projektets informanter (workshop- og udvalgsmedlemmer) indbudt til at deltage i en afsluttende spørgeskemaundersøgelse der fokuserede på en række forskellige samfundsmæssige aspekter af sprogteknologien: handel, arbejdsmarked, serviceniveau i den offentlige sektor og innovation.

Vækstpotentialet for dansk sprogteknologi i de kommende år



Figur 1. Grafen samler alle svar afgivet ved den afsluttende spørgeskemaundersøgelse.

Undersøgelsen viste den samme tendens inden for alle aspekter. Samtlige interessegrupper forventer en markant stimulerende virkning i alternativscenariet, både inden for økonomi, velfærd og innovation. Som grafen i figur 1 viser, vurderes sandsynligheden for vækst i de kommende år som "meget sandsynligt" i alternativscenariet, men "kun lidt sandsynligt" i basisscenariet af det store flertal af informanter. De to optimistiske vurderinger ("meget sandsynligt" og "vil helt sikkert ske") har tre gange større tilslutning i alternativscenariet end i basisscenariet. De to pessimistiske vurderinger ("ikke sandsynligt" og "kun lidt sandsynligt") har tre gange større tilslutning i basisscenariet end i alternativscenariet. Se datagrundlaget i bilag 1.

Markedets aktuelle værdi

En realistisk vurdering af det samlede sprogteknologiske marked i Danmark bør efter vores opfattelse tage udgangspunkt i det markedssegment som kan kvantificeres mest pålideligt, og det vil i praksis sige den del der er baseret på licensstruktur. Denne del udgør 30-50 % af det samlede marked for sprogteknologi ifølge leverandørernes og slutbrugerrepræsentanternes feedback (data fra spørgeskemaundersøgelser, workshops og personinterviews).

Licensmarkedet udgøres hovedsageligt af talegenkendelsesløsninger til den offentlige sektor (kommuner, regioner og staten) markedsført af de fem største leverandørfirmaer (Nuance, KMD, MaxManus, Dictus og Mirsk). Vi har opfordret både disse leverandører og en række af deres købere til at vurdere licensmarkedet i Danmark (både antal personlicenser og prisstruktur). Ikke alle har ønsket at bidrage, og andre har betinget sig fortrolighed. Tallene herunder er derfor medianværdier af alle de indkomne data.

Aktuelle licensaftaler 2019 (talegenkendelse, fuldtidsansatte brugere)	**14.000 årslicenser**
Gennemsnitlig pris per årslicens	**DKK 9.000**

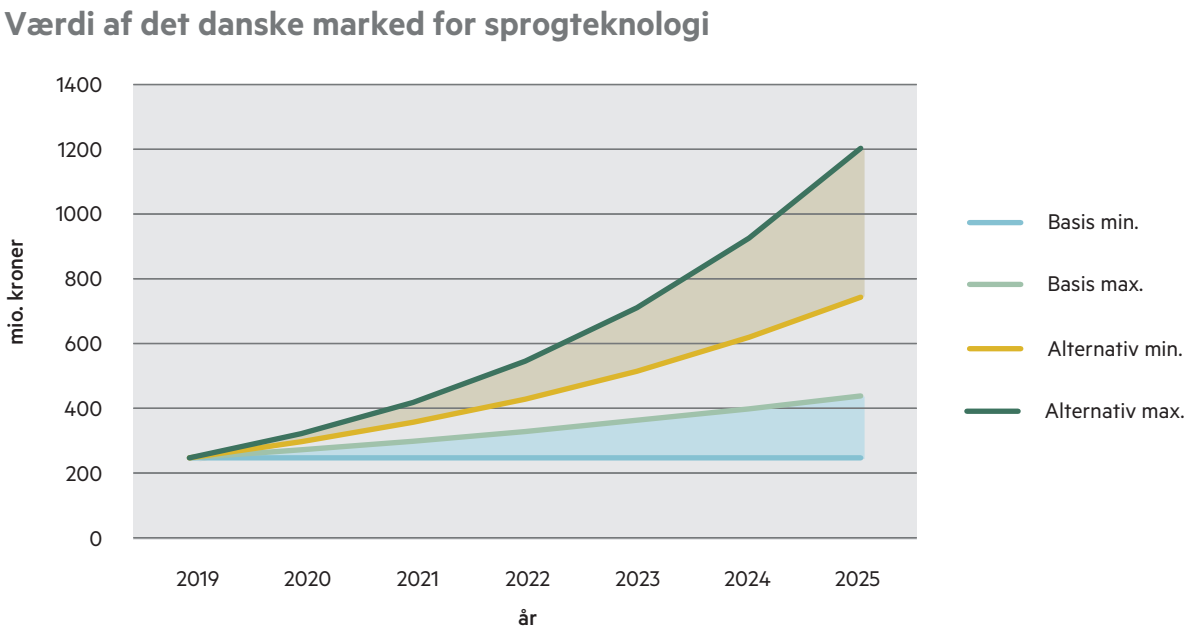
Baseret på disse tal udgør den licensbaserede del af dansk sprogteknologi som minimum DKK 126 mio. om året. Dette skøn er konservativt, da det kun er baseret på kendte tal, samt kun på én enkelt teknologisk sektor, dvs. ikke inddrager licensstrukturen inden for fx. medieservice, tekstanalyse og oversættelse (disse sektorer er ikke organiseret på tværs af brugergrupperne og er derfor langt sværere at kvantificere). Da licensmarkedet, som før nævnt, kan antages højst at udgøre 50 % af den samlede markedsværdi, kan vi hermed uddrage et estimat for hele markedet på mindst DKK 250 mio. årligt.

Fremskrivning

I spørgeskemaundersøgelsen blev deltagerne bedt om at kommentere dette scenarie: "Udviklingen inden for dansk sprogteknologi vil vokse støt (med samme eller større vækstrate) i de kommende år". De to optimistiske vurderinger ("meget sandsynligt" og "vil helt sikkert ske") stod for hhv. 29 % i basisscenariet og 95 % i alternativscenariet. Topvurderingen alene fik ingen bidrag i basisscenariet, men 40 % i alternativscenariet.

Vi konkluderer at branchen forventer ingen eller ringe vækst i basisscenariet, mens den i alternativscenariet forventer en betydelig vækst. I statistiske termer ventes, som minimum, en eksponentiel udvikling i de kommende år (vækst med fast vækstrate). Den forventede vækstrate fremgår ikke af spørgeskemaundersøgelsen, men ved personlige interviews i interessegruppen Leverandører nævnes årlige vækstrater fra 20-50% som mest sandsynlige i alternativscenariet, med medianen 25 %. Denne værdi er naturligvis behæftet med betydelig usikkerhed, men virker ikke urealistisk i lyset af den generelle optimisme omkring virkningerne af de anbefalede tiltag.

Givet en vækstrate på mellem 20 % og 30 % i alternativscenariet og mellem 0 % og 10 % i basisscenariet kan værdien af det sprogteknologiske marked fremskrives som vist i figur 2.



Figur 2. Den forventede vækst i dansk sprogteknologi i hhv. basisscenariet og alternativscenariet

Samlet vurdering

Værdien af det aktuelle danske sprogteknologiske marked udgør, i en konservativ vurdering, DKK 250 mio. om året. Den reelle værdi kan formodes at ligge 25-50 % højere. Hvis de anbefalede tiltag effektueres, forventer alle relevante interessegrupper at se en markant vækst i dansksprogteknologi inden for handel, industri og offentlig service (alternativscenariet). I modsat fald forventes en stagnation på alle de nævnte områder (basisscenariet).

På baggrund af de indsamlede data vurderes effekten af de anbefalede tiltag til omkring 20 procentpoint i øget årligt vækstrate, svarende til en værdistigning på mindst DKK 50 mio. kroner i 2020, voksende til mindst DKK 500 mio. i 2025. Til sammenligning udgør de anbefalede årlige statstilskud ikke over DKK 20 mio. kroner årligt i samme periode.

På denne baggrund konkluderer vi at de anbefalede aktiviteter med de foreslåede økonomiske rammer udgør en attraktiv samfundsmæssig business case.


DANSK
ALLE HAR BRUG FOR SPROGTEKNOLOGI
SPROGTEK2018.DK

10. Konklusion

Vi har fremlagt en række anbefalinger til offentligt støttede aktiviteter som, tilsammen eller enkeltvis, vil have en stimulerende virkning på fremstillingen og anvendelsen af dansk sprogteknologi i mange år fremover. Anbefalingerne er resultatet af en omfattende vidensindsamling med bidrag fra alle de største professionelle interessegrupper i Danmark inden for sprogteknologi. Vi har prioriteret anbefalingerne sådan at de mest påkrævede aktiviteter kan iværksættes først, men vi understreger at de fire overordnede indsatsområder er indbyrdes afhængige og derfor bør sættes i gang samtidig.

Særligt central er fremstillingen og distributionen af de nøglekomponenter som p.t. mangler for det danske sprog, og som især de små og mellemstore aktører i markedet ikke har mulighed for selv at fremstille. Uden disse nøglekomponenter vil dansk sprogteknologi i fremtiden formentlig kun blive fremstillet og markedsført af store, internationale leverandører der ikke nødvendigvis ser det danske sprog som værdifuldt i sig selv. De seneste 20 år har i høj grad været præget af en udvikling i denne retning. Man kan ikke forvente at disse leverandører vil sikre det danske sprogs integritet i en tid hvor den mest lukrative del af markedet (fx shipping, medicinsk og kemisk industri, robotindustrien, banksektoren, rumfart, energiforsyning, computerspilsindustrien, m.m.) i forvejen er fristet til helt at gå væk fra dansk som forretningssprog, selv i kommunikationen mellem danskere og skandinaver (fx i firmaer, værksteder, laboratorier og på uddannelser).

Den første og vigtigste aktivitet angår derfor oprettelsen af en dansk sprogbank som kan koordinere udviklingen og distributionen af danske sprogkomponenter. Ved at fremstille ord- og termdatabaser, tekstsamlinger, opmærkede guldstandarder og talematerialer i højeste lingvistiske kvalitet og gøre dem frit tilgængelige for både private borgere, små og store virksomheder og offentlige organisationer vil staten styrke dansk sprogteknologi og derigennem sikre at vores sprog, også for de næste generationer, er brugbart i enhver sammenhæng.

10.1. Udvalgets svar på de spørgsmål som blev stillet i kommissoriet

Inden for hvilke sektorer og erhverv vil der i de kommende 10 år være størst behov for digitale tjenester og applikationer baseret på kunstig intelligens på dansk og andre sprog?

De digitale tjenester som der vil være mest behov for er

1. Talebaserede grænseflader (dialogsystemer) og digitale assistenter
2. Systemer til analyse af teksters indhold og klassifikation af tekster
3. Systemer og metoder til håndtering af terminologi.

Behovet for talebaserede grænseflader vil primært være inden for

1. Automatisering af tasteprocesser, fx inden for sundhedssektoren og offentlig service
2. Automatisering af telefonbetjening, fx inden for offentlige virksomheder og private virksomheder som banker, pensionsselskaber mfl.
3. Automatisering af styreprocesser, fx af biler, robotter og maskiner generelt
4. Automatisering af dialog, fx via chatrobotter.

Behovet for systemer til analyse af teksters indhold og klassifikation af tekster vil primært være inden for

1. Teksttunge offentlige institutioner, som fx domstole, ministerier og styrelser, kommuner og regioner, politi, militær, samt statslige virksomheder som fx ATP
2. Offentlige og private virksomheder med mange kundehenvendelser via e-mail og andre skriftlige kanaler
3. Teksttunge private institutioner, som fx advokatvirksomheder, mediebranchen, teleindustrien, pensionsbranchen, banker, produktionsvirksomheder med store dokumentationskrav, oversættelsesvirksomheder, m.fl.

Behovet for systemer og metoder til håndtering af terminologi vil primært være inden for

1. Statslige styrelser og kommuner samt sundhedssektoren bl.a. i forbindelse med implementering af digitaliseringsklar lovgivning og kvalitetssikring af processer, it-systemer mv.
2. Virksomheder med høje dokumentationskrav, fx banker, pensionsselskaber, medicinalindustrien, advokatvirksomheder, it-virksomheder m.fl.

Hvilke udfordringer ser virksomheder og offentlige institutioner i forhold til at udvikle disse tjenester og applikationer – og hvilke udfordringer bliver overset?

Udfordringer:

1. At der ikke findes systemer i tilstrækkelig høj kvalitet der kan håndtere dansk
2. At udenlandske aktører ikke betragter dansk som et attraktivt marked
3. At initialomkostningerne er for store til at de kan løftes af den enkelte virksomhed
4. At der ikke findes tilstrækkelig viden om hvordan sprog og kunstig intelligens interagerer.

Det bliver ofte overset at indføring af sprogteknologi og kunstig intelligens er mindst lige så krævende som indførelse af andre former for ny teknologi, og derfor afsættes der ikke tiltrækkeligt med resurser til at inddrage og opkvalificere medarbejderne hvilket resulterer i at systemerne ikke udnyttes optimalt.

På hvilken måde kan sprogteknologi bidrage til at sikre en bedre og billigere offentlig service?

Billigere offentlig service:

1. Større effektivitet og bedre multi-tasking gennem taleinput i stedet for tasteinput
2. Større effektivitet ved behandling af borgerhenvendelser
3. Større effektivitet ved analyse af indberetninger fra virksomheder og borgere
4. Større effektivitet ved udvikling af digitaliseringsklar lovgivning og it-systemer.

Bedre offentlig service:

1. Hurtigere betjening af borgere
2. Flere muligheder for selvbetjening
3. Bedre muligheder for at tilgå information fra det offentlige
4. Bedre service for borgere med kommunikationsvanskeligheder
5. Mere pålidelige it-systemer og dermed større sikkerhed for den enkelte
6. Bedre service for mennesker som taler andre sprog.

På hvilken måde kan erfaringer fra andre lande, EU og Norden nyttiggøres?

1. Både i EU og i andre nordiske lande har man i langt højere grad end i Danmark arbejdet med at styrke sprogteknologi. Erfaringerne viser at en kontinuerlig strategisk indsats, som det fx er sket i Nederlandene og Sverige, øger landenes muligheder for at udvikle sprogteknologi i høj kvalitet for deres sprog, og at det tilskynder både udenlandske aktører og hjemlige virksomheder til at investere i at udvikle tjenester for det pågældende sprog. Endvidere viser erfaringerne at indsamling af data i sig selv ikke har nogen stor effekt, hvorimod et tæt og målrettet samarbejde mellem de ansvarlige for dataindsamling og de relevante aktører i markedet samt med centrale forskningsinitiativer er en mere lovende strategi.
2. Der kan i høj grad drages paralleller til EU og Norden, og der kan opnås synergieffekter gennem et tæt samarbejde. Danmark kan drage nytte af EU's arbejde med det digitale indre marked, herunder fjernelsen af de sproglige barrierer, og af de planer der ligger for kunstig intelligens og sprog i de nye Horizon Europe-programmer. Der er et stort beredskab til samarbejde både mellem sprogmiljøerne og mellem digitaliseringsstyrelserne i de nordiske lande, men der har i hvert fald for sprogmiljøernes vedkommende ikke været afsat tilstrækkeligt med resurser til at høste de oplagte gevinster et nordisk samarbejde om sprogteknologi kunne give, bl.a. i kraft af ligheden mellem de skandinaviske sprog.

Hvilke vækst- og jobmuligheder ligger der i en satsning på dansk sprogteknologi?

1. Alle virksomheder som arbejder med sprogteknologi, mærker en tydelig stigning i efterspørgslen
2. Samtidig ses en lang række forsøg i virksomheder og offentlige institutioner på at afprøve metoder der involverer kunstig intelligens, på interne data – typisk med værktøjer som er beregnet til engelske data. Mange af disse forsøg strander på et manglende kendskab til de udfordringer som det danske sprog giver
3. Der er et godt vækstgrundlag for etablerede sprogteknologivirksomheder og for virksomheder som kan yde konsulentbistand til udvikling af nye analysemetoder og andre tjenester. Endvidere er det sandsynligt at udenlandske virksomheder vil efterspørge dansk arbejdskraft til tilpasning af deres produkter til dansk
4. Der ligger permanente jobs i løbende at opdatere systemerne og tilpasse dem ikke blot den teknologiske, men også den sproglige udvikling, fx generelle ændringer i sprogbrugen, ændring af lovtekster eller betegnelser for ydelser (fx integrationsydelse -> hjemsendelsesydelse).

Hvad er den samfundsøkonomiske business case set i forhold til investeringsbehovet?

Investeringsbehovet er ca. 14-15 mio.kr om året over en 4-årig periode. Derudover bør der foretages en langsigtet investering i forskning og uddannelse.

1. Det er naturligvis vanskeligt at gøre op i kroner og øre, da investeringen både gør sig gældende som effektivisering og som en bedre serviceoplevelse hos borgerne. På den anden side vil bedre service formentlig medføre færre henvendelser og færre misforståelser og dermed også tidsbesparelser i den offentlige sektor
2. En massiv gevinst vil kunne ses ved samspillet mellem begrebsarbejde, terminologi og it-systemer hvor bedre forståelse af sprogets betydning, større databevidsthed og bedre datadisciplin vil kunne forhindre at store it-systemer skal skrottes fordi de viser sig at være uanvendelige
3. Se endvidere kapitel 9.2.

Hvilke politiske tiltag kan foreslås for at understøtte virksomheder og offentlige institutioner i at inddrage dansk og andre sprog når der skal udvikles og anvendes nye teknologier baseret på kunstig intelligens?

Se anbefalingerne.

Hvilken betydning får en satsning på dansk sprogteknologi for udviklingen af det danske sprog, for samfundets udvikling og for den enkelte?

Betydningen for det danske sprog:

Det er et helt grundlæggende vilkår for al sproglig udvikling at sproget bliver brugt i alle samfundets områder. Hvis sproget ikke bruges, mister det sin evne til at udtrykke det som vi har behov for, på en smidig og hensigtsmæssig måde. Det danske sprog i sig selv er ikke umiddelbart truet hvis vi ikke får adgang til god sprogteknologi på dansk, men der vil være et stigende antal sammenhænge hvor danskerne enten skal stille sig tilfredse med dårlig og fejlagtig sprogbrug, misforståelser og fejl fra systemernes side eller skifte over til at bruge engelsk. Sammen med den stigende brug af engelsk på universiteter og i erhvervslivet kan dette på længere sigt føre til en lavere funktionsdygtighed på disse områder. Udvalget ser snarere de største effekter og gevinster i forhold til de mennesker der bruger sproget frem for i forhold til selve sproget, og udvalgets anbefalinger fokuserer derfor udelukkende på de samfundsmæssige konsekvenser og konsekvenserne for den enkelte.

Samfundsmæssige konsekvenser:

1. Udvikling af sprogteknologi for dansk sikrer at offentlige og private digitale tjenester der bruger sprog som kommunikationsmiddel, kan servicere mennesker som bruger det danske sprog
2. Udvikling af sprogteknologi for dansk sikrer endvidere at danske data – både nutidige og historiske, almensproglige og fagsproglige – kan inddrages i systemer som benytter sig af kunstig intelligens, således at de kan afspejle ikke blot viden om dansk sprog, men også om danske kultur- og samfundsforhold

3. Hvis der ikke er god adgang til danske data og sprogteknologiske værktøjer der kan håndtere dansk, vil offentlige og private tjenester der skal servicere danske borgere og medarbejdere i danske virksomheder, enten levere en ringere kvalitet eller slet ikke levere tjenesten på dansk, men i stedet på engelsk
4. Hvis der ikke er god adgang til danske data og sprogteknologiske værktøjer der kan håndtere dansk, vil systemer som baseres på kunstig intelligens, få mindre relevans og nytte for samfundet, da de ikke i samme omfang vil kunne afspejle danske kultur- og samfundsforhold
5. Hvis der ikke er god adgang til danske data og sprogteknologiske værktøjer der kan håndtere dansk, risikeren man at borgerne og arbejdsstyrken ikke tager imod de muligheder og effektiviseringsgevinster den nye teknologi tilbyder, fordi de simpelthen afviser at benytte sig af den
6. Hvis der ikke er god adgang til danske data og sprogteknologiske værktøjer der kan håndtere dansk, vil Danmark få vanskeligere ved at høste de gevinster der ligger i EU's digitale indre marked, da det vil blive vanskeligere at overvinde de sprogbarrierer som EU satser på at fjerne i samarbejde med medlemslandene.

Konsekvenser for den enkelte:

1. Hvis der ikke er god adgang til danske data og sprogteknologiske værktøjer der kan håndtere dansk, vil mennesker med behov for kommunkationshjælpemidler være ringere stillet i dagligdagen og især have mindre mulighed for at deltage i samfundet samfundet på lige fod med andre
2. Hvis der ikke er god adgang til danske data og sprogteknologiske værktøjer der kan håndtere dansk, vil mennesker som ikke har dansk som modersmål have sværere ved at deltage i samfundet
3. Hvis der ikke er god adgang til danske data og sprogteknologiske værktøjer der kan håndtere dansk, vil det enkelte menneskes mulighed for at benytte sig af ny teknologi blive forringet.

Hvordan sikres det at der udvikles dansksproget sprogteknologi?

Se anbefalingerne.

Hvad er fordelene og ulemperne ved udvikling af dansk sprogteknologi i Danmark?

Fordele:

1. At der bliver udviklet sprogteknologi for dansk også i de tilfælde hvor udenlandske virksomheder ikke har et incitament til at investere
2. At udviklingen sker i tilnærmelsesvist samme tempo som for fx engelsk og andre hovedsprog og ikke som det er tilfældet i dag med 4-5 års forsinkelse
3. At systemerne opnår en kvalitet som er tilstrækkelig god til at brugerne vil anvende dem
4. At vi bestemmer over vores egne sprogresurser
5. At borgere får adgang til service på deres eget sprog
6. At vi kan indhøste de effektiviseringsgevinster som teknologien tilbyder
7. At vi kan tilbyde forskere og udviklere danske data og danske resurser til at videreudvikle systemerne i fremtiden.

Ulemper:

1. At der er behov for investeringer i en sprogteknologisk infrastruktur
2. At den sprogteknologiske infrastruktur løbende skal vedligeholdes.

Hvordan kan det sikres at der fortsat uddannes mennesker med tilstrækkelige kompetencer inden for dansk sprogteknologi?

Se anbefalingerne.

Hvilket behov er der for udvikling af en dansk termbank, hvilke domæner skal den dække, og hvordan kan den bedst gøres tilgængelig?

Se anbefalingerne i afsnit 8.2.4 og generelt om terminologi i afsnit 2.5.

Bilag

Oversigt over bilag:

Bilag 1

Data fra spørgeskemaundersøgelser i forbindelse med udvalgets workshops og seminarer

Bilag 2

Referat af workshop om terminologi

Bilag 3

Referat af workshop om automatisk oversættelse

Bilag 4

Oversigt over institutioner og virksomheder der har bidraget til udvalgets arbejde

BILAG 1

I dette bilag fremlægges data fra spørgeskemaundersøgelserne i forbindelse med workshops og seminarer.

- Spørgeskema for alle deltagere (udsendt i forbindelse med det afsluttende seminar)
- Spørgeskema for slutbrugere
- Spørgeskema for leverandører
- Spørgeskema for udviklere
- Spørgeskema for forskere og formidlere

Spørgeskema for alle deltagere

Første undersøgelse - Basisscenariet

Forud for slutseminaret blev alle projektets deltagere (workshopdeltagere og udvalgsmedlemmer) bedt om at estimere den fremtidige udvikling for dansk sprogteknologi. Estimatet skulle tage udgangspunkt i basisscenariet (dvs. antage at ingen af Sprogteknologiudvalgets anbefalinger følges). Undersøgelsen fokuserede på otte forskellige samfundsrelevante aspekter (se venstre kolonne i tabellen).

Anden undersøgelse - Alternativscenariet

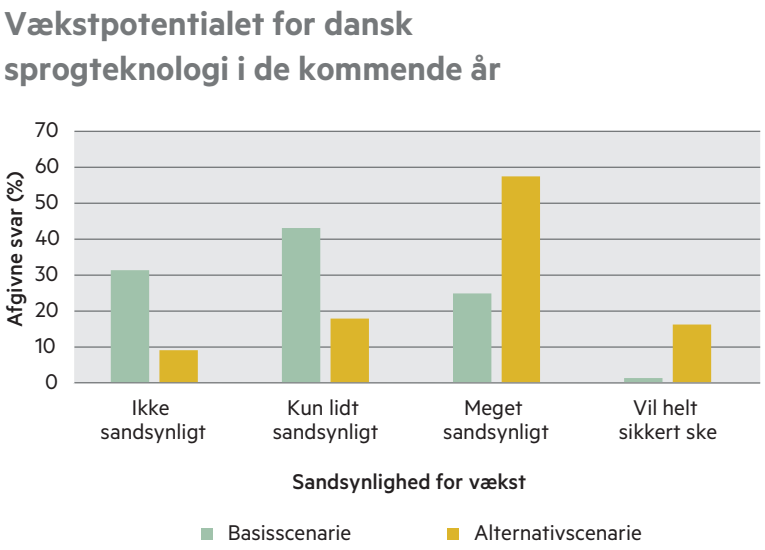
Umiddelbart efter slutseminaret blev alle projektets deltagere igen bedt om at estimere den fremtidige udvikling for dansk sprogteknologi i de samme otte aspekter, dog denne gang i Alternativscenariet (hvor alle Sprogteknologiudvalgets anbefalinger følges).

Besvarelser

Besvarelserne er samlet i tabellen herunder og i blokdiagrammet figur 1.

Aspekter af dansk sprogteknologi	Basisscenarie					Alternativscenarie				
	ikke sand-syn-ligt	kun lidt sand-syn-ligt	meget sand-syn-ligt	vil helt sik-kert ske	VED IKKE	ikke sand-syn-ligt	kun lidt sand-syn-ligt	meget sand-syn-ligt	vil helt sik-kert ske	VED IKKE
1. Udviklingen inden for dansk sprogteknologi vil vokse støt (med samme eller større vækstrate) i de kommende år	9	8	7	0	1	0	1	15	11	0
2. I fremtiden vil størstedelen af leverancerne af dansk sprogteknologi komme fra udenlandske leverandører	0	2	20	2	1	2	11	6	0	8
3. En underskov af nye iværksættere gør Danmark førende på nogle applikationsområder (fx services i den offentlige sektor, udvikling af nye processer i industrien)	8	11	1	0	5	0	2	19	1	5
4. Udviklingen inden for dansk sprogteknologi vil medføre øget ledighed i de berørte sektorer af det danske arbejdsmarked	3	2	8	0	12	15	10	0	0	2
5. Udviklingen inden for dansk sprogteknologi giver forøget beskæftigelse i de berørte sektorer af det danske arbejdsmarked	11	7	2	0	5	0	5	18	1	3
6. Udviklingen inden for dansk sprogteknologi giver forbedrede velfærdstilbud til danske borgere (fx bedre sundhed og bedre hjælpemidler)	6	16	2	0	1	0	2	16	8	1
7. Udviklingen inden for dansk sprogteknologi giver højere økonomiske vækstrater inden for de berørte erhverv	9	13	0	0	3	0	3	17	3	4
8. Udviklingen inden for dansk sprogteknologi giver øget innovation i den private og den offentlige sektor	7	14	2	0	2	0	0	19	7	1

Data fra den afsluttende spørgeskemaundersøgelse for alle interessegrupper



Figur 1. Samlet vurdering af vækstpotentialet for dansk sprogteknologi, baseret på de summerede besvarelser fra aspekterne 1-8 (jf. tabellen). De to negative aspekter 2 og 4 bidrager med faktor -1.

Spørgeskema for slutbrugere

Spørgsmålslisten

1) Hvilke former for sprogteknologi bruges i din organisation?
2) Hvordan bruger du selv sprogteknologi til dine daglige arbejdsopgaver?
3) Hvor har du mest nytte af sprogteknologien? (Fx mht. arbejdsglæde og produktivitet, samt afhjælpning af fysiologiske eller sproglige udfordringer).
4) Hvilke af dine arbejdsområder kunne sprogteknologien støtte bedre? Giv gerne konkrete eksempler.
5) Er dit områdes fagtermer tilstrækkeligt understøttet? Hvis nej, hvilke ord og vendinger volder problemer?
6) Fungerer sprogteknologien optimalt rent teknisk? Er den fx let at starte og let at anvende?
7) Hvis du oplever besvær, hvad forhindrer teknologien i at fungere optimalt i din arbejdssituation?
8) Hvordan bliver du inddraget når der indføres ny sprogteknologi? Bliver du tilstrækkeligt informeret og efteruddannet?
9) Hvor væsentligt er det at teknologien fungerer på dansk? Kunne du overveje at skifte til engelsk?
10) Hvilke fremtidige anvendelser af dansk sprogteknologi ser I i din organisation? (om 1 år og om 5 år)

Besvarelser

Besvarelserne blev givet under fortrolighedsløfte. Her fremlægges de tendenser der har relevans for udredningsarbejdet.

	Spørgsmål	Resumerede besvarelser
1	Hvilke sprogteknologier bruger din organisation?	*Organisationens vigtigste anvendte sprogteknologi*: Talegenkendelse: 6, oversættelse: 5 *Alle organisationens anvendte sprogteknologier*: Talegenkendelse: 6, oversættelse: 6, tekstanalyse: 4, information extraction: 3, tekstklassifikation: 2, talesyntese: 1, elektroniske ordbøger: 1
2	Hvordan bruger du selv sprogteknologi?	Jeg benytter sprogteknologisk værktøj dagligt: 5 Jeg rådgiver organisationens medarbejdere: 3 Jeg rådgiver brugere uden for min organisation: 2 Jeg udvikler løsninger der bygger på sprogteknologi: 2
3	Hvor har du mest nytte af sprogteknologi?	Højere kvalitet (produkter/ydelser): 8 Større produktivitet (produkter/ydelser): 7 Bedre arbejdsvilkår (afhjælper handikap og udfordringer): 1 Sprogteknologi er organisationens kerneydelse: 1
4	Hvor kunne sprogteknologien støtte bedre?	Værktøj med bedre sproglig ydelse: 5 Værktøj med bedre teknologisk integration: 3 Bedre deling af resurser: 3 Intet behov for forbedring: 1
5	Er fagtermer tilstrækkeligt understøttet?	*Svarene kredser om personlige og fagspecifikke detaljer som er belyst andetsteds i rapporten (se afsnit om terminologi)*
6, 7, 8	Tekniske aspekter (anvendelse og indføring af programmel)	Ingen problemer: 5 Problemer med integration (formater, software, OS): 4 Udfordring med træning af brugere: 2
9	Hvor væsentligt er det at teknologien fungerer på dansk?	*Spørgsmålet opfattes forskelligt: (i) om teknologien kan behandle dansk sprog, (ii) om brugergrænsefladen er på dansk. Alle 6 svar i første kategori er:* "afgørende vigtigt"
10	Hvilke fremtidige anvendelser ser I?	Bedre talegenkendelse (fx til 'almindeligt' dansk): 6 Nye kombinationer af sprogteknologier og AI: 3 Automatisk oversættelse til og fra mange nye sprog: 2 *(svar om firmaspecifikke produkter/ydelser er ikke medregnet)*

Spørgeskema for leverandører

Spørgsmålslisten

1) Hvilke sprogteknologiske produkter udbyder I?
2) Hvilke sprogteknologiske moduler af jeres software har I selv udviklet?
3) Hvem er jeres kunder?
4) Hvor mange slutbrugere har I?
5) Sælger I licenser? Hvordan er licensstrukturen?
6) Efterspørgslen på jeres sprogteknologiske produkter? Opleves stigning/fald?
7) Hvad er en god business case for jer?
8) Hvad er et optimalt udbudsscenarie for jer?
9) Hvordan afdækker I brugernes behov?
10) Tager I på forhånd hensyn til det fagområde som applikationen skal fungere i?
11) Hvordan interagerer I med brugerne før en leverance?
12) Hvordan forbereder I en leverance?
13) Hvordan integreres jeres sprogteknologiske produkter i kundens andre produkter/systemer?
14) Hvordan interagerer I med brugerne når produktet er indfaset?
15) Hvordan inddrager I brugerne i forbedringen af systemet?
16) Hvordan tager I hensyn til arbejdspladsens særlige termer og sprogbrug?
17) Hvordan kan kunderne indberette fejl og komme med forslag til forbedringer?
18) Oplever I behov hos kunderne som I ikke kan dække? Hvilke? Hvad er barriererne?
19) Hvor meget nytte har I af kundernes feedback?
20) Hvordan ser I konkurrencesituationen på markedet for dansk sprogteknologi?
21) Hvordan samarbejder I med andre udbydere, leverandører etc.?
22) Kan I se fordele i at dele sprogteknologiske ressourcer med andre virksomheder?
23) Kan I se fordele i at dele sprogteknologiske ressourcer på open source-basis?
24) Hvilke behov for sprogteknologiske byggesten eller moduler ser I i fremtiden?
25) Hvilke brancher ser I som interessante ift. fremtidig brug af sprogteknologi?
26) Hvordan ser I forholdet mellem dansk og engelsk sprogteknologi nu og i fremtiden?
27) Hvordan forholder I jer til Den Fællesoffentlige Digitaliseringsstrategi?
28) Hvilke nationale tiltag kunne forbedre markedet for jeres produkter?

Besvarelser

Besvarelserne af spørgeskemaet blev givet under fortrolighedsløfte.

	Spørgsmål	Resumerede besvarelser
1	Hvilke produkter udbyder I?	*Firmaets vigtigste produkt:* Talegenkendelse: 4 - tekstanalyse: 2 - information extraction: 2 - talesyntese: 1 - elektroniske ordbøger: 1 *Alle firmaets produkter:* Tekstanalyse: 5 - talegenkendelse: 4 - information extraction: 3 - elektroniske ordbøger: 2 - maskinoversættelse: 1 - talesyntese: 1
2	Hvilke moduler har I selv udviklet?	Alle: 6 - Mange: 4 - Visse: 1 - Ingen: 1
3	Hvem er jeres kunder?	*Sektorer (én per besvarelse):* Offentlige: 6 - private: 1 - offentlige & private: 5 *Organisationer (gerne flere per besvarelse):* Kommune: 10 - Stat & region: 6 - Store virksomheder: 5 - Andre: 4
4-5	Antal slutbrugere? Hvilken licensstruktur?	*(en del besvarelser kan ikke citeres)* *Antal slutbrugere:* Tal mellem 400 og 800.000 angives. *Licensbaseret?* Ja: 13 - Nej: 0 - Licens & produktsalg: enkelte
6	Oplever I stigende efterspørgsel?	11 ud af 13 svarer "stigende efterspørgsel". Ingen konstaterer fald. *To uddyber:* "forskellig for forskellige produkter", "Vi oplever efterspørgsel på nye, intelligente måder at bruge vores eksisterende softwareløsninger på. Flere handler om inddragelse af NLP."
7-8	Optimale betingelser for markedsføring	*Typiske svar:* "mulighed for dialog inden udbud", "udbud som matcher firmaets produktportefølje", "god behovsafklaring hos kunden", "god intern kommunikation i kundens organisation", "salg med udviklingsfinansiering", "langsigtede kontrakter"
10, 11, 12, 14, 15, 17, 19	Interaktion med kunden	*Typiske svar:* "Design Thinking og bruger-interviews", "dialog og demonstrationer", "arbejdsgangsanalyser hos kunden", "løbende statusmøder og workshops", "hotline (email, telefon)", "fortsatte undervisningsforløb og kurser", "observationer af kundens adfærd", "brugerinterviews"
13	Integration med kundens environment	*Typiske svar:* "Integration via styresystemer og plugins", "API", "åbne standarder, HTTPS, JSON, SMTP", "fokus på at vores produkter overholder dokumentudvekslingsstandarder"
16	Hvordan tager I hensyn til arbejdspladsens særlige sprogbrug?	*Typiske svar:* "møder med kunden", "tekstindsamling hos kunden", "fagspecifikke ordlister og ordbøger", "sprogmodeller udvikles specifikt til området", "specifikke termer integreres i maskinlæring"
18	Kundebehov I ikke kan dække?	*Typiske svar:* "problemer med forældede fagsystemer", "talegenkendelse inden for helt nye fagområder", "urealistiske forventninger hos kunden"
20-21	Samarbejde med andre leverandører	Vi samarbejder ikke med andre leverandører: 5 Vi samarbejder med kundernes øvrige leverandører: 3 Vi samarbejder med andre leverandører: 3 Vi leverer til andre leverandører: 1
22	Fordele ved at dele resurser med andre virksomheder?	Ja: 5 Betinget ja: 5 Neutral: 0 Betinget nej: 1 Nej: 1

23	Fordele ved at dele resurser open-source?	Ja: 3 Betinget ja: 1 Neutral: 2 Betinget nej: 1 Nej: 3
24, 18	Fremtidige behov for nye sprogresurser	Frie korpusser til træning af talegenkendelse: 5 Frie korpusser til træning af tekstanalyse: 3 Fri ordbase (annoteret for morfologi, semantik, udtale osv.): 6 Andre typer fri resurser og værktøj: 6
25	Hvilke brancher ser I som interessante ift. fremtidig brug af sprogteknologi?	*Mange svarer:* "alle brancher", "sprogteknologi er ikke branchespecifik" o.lign. *Konkrete brancher nævnt:* finans, jura, retail, undervisning, professionel rådgivning (skrift og tale), forsikring, pension, videnskabelig undersøgelse og forskning, salg og marketing, produktudvikling, kommunikationsstøtte til mennesker med særlige behov
26	Hvordan ser I forholdet mellem dansk og engelsk sprogteknologi nu og i fremtiden?	*Typiske svar:* "Kvaliteten er langt bedre for engelsk", "P.t. er primært udvikling indenfor engelsk", "Der er stort behov for dansk sprogteknologi indenfor vores kundesegment", "Der er en kæmpe forskel på præcision, som det bliver svært at gøre noget ved uden ordentlige datasæt", "mindre konkurrence på det danske marked", "engelsk sprogteknologi spiller stadig en væsentlig rolle blandt danske forbrugere", "afgørende at have data og værktøj, der behandler dansk for at kunne tilbyde intellligente løsninger der inddrager sprogteknologi, til danske kunder", "Vi tror på, at folk stadig gerne vil kunne tale dansk til deres applikationer"
27	Hvordan forholder I jer til den Fællesoffentlige Digitaliseringsstrategi?	*Typiske svar:* "Følger den, når den er relevant", "Vores produkter understøtter den direkte", "Vi ser os selv som en vital part af Danmarks digitalisering", "Uambitiøs", "Vi oplever ikke at virksomheder af vores størrelse - indtil nu - har fået hjælp til at gøre os relevante som udbydere" [besvarelser fra en mindre virksomhed].
28	Hvilke nationale tiltag kunne forbedre markedet for jeres produkter?	*Typiske svar:* "Adgang til træningsmaterialer og transkriptionsdata", "Krav om dansk i stedet for engelsk", "Oplysning om ordblindhed", "Investering i dansk sprogteknologi og i projekter hvor dansk sprogteknologi kan bidrage", "Værktøjer til semantisk analyse og terminologi", "Fælles ordbase af høj kvalitet", "Fælles regler for datamodellering", "Tættere samarbejde med offentlige myndigheder"

Spørgeskema for udviklere

Spørgsmålslisten

1) Hvilket firma repræsenterer du?
2) Hvilke sprogteknologiske produkter udvikler I?
3) Arbejder du med fagsprog?
4) Arbejder du med alment sprog?
5) Hvilken/hvilke typer af sprogteknologisk software udvikler du?
6) Hvilke sprog arbejder du med?
7) Hvilken/hvilke typer af sprogteknologisk software udvikler din organisation?
8) Hvilken/hvilke softwareplatforme bruger du?
9) Hvilke open-source sprogteknologiske værktøjer for dansk kender du?
10) Hvilke af dem bruger du?
11) Hvilke af dem bruger du ikke? Hvorfor?
12) Bruger du sprogneutrale værktøjer/softwarepakker? Hvilke?
13) Bruger du sprogspecifikke værktøjer/softwarepakker? Hvilke?
14) Hvordan træffer du valg af softwareplatform i dit udviklingsarbejde?
15) Ville du træffe det samme valg af softwareplatform hvis du skulle begynde nu?
16) Hvilke metoder bruger du?
17) Ville du bruge de samme metoder hvis du selv kunne bestemme?
18) Hvad oplever du som den største teknologiske begrænsning?
19) Hvilke danske sproglige data arbejder du med?
20) Hvilke sproglige data for dansk mangler du?
21) Hvilke sproglige data ville du have mest nytte af at kunne få fra en offentlig sprogbank?
22) Hvis du skulle prioritere én resurse, hvad skulle det så være?
23) Ville det rent teknologisk være en fordel at dele resurser med andre firmaer?
24) Ville det rent teknologisk være en fordel at dele resurser på open-source-basis?

Besvarelser

Besvarelserne blev givet under fortrolighedsløfte.

	Spørgsmål	Resumerede besvarelser
1	Hvilket firma repræsenterer du?	*(besvarelserne er anonyme)*
2, 5, 7	Sprogteknologiske produkter	*Organisationens vigtigste udviklingsområde*: Tekstklassifikation: 4 Taleteknologi: 4 Maskinoversættelse: 2 *Alle organisationens udviklingsområder*: Taleteknologi: 7 Tekstklassifikation: 5 Information extraction: 5 Maskinoversættelse: 2 *Produktporteføljer:* Hjælpemidler, læremidler, talesyntese, grammatikværktøjer, tekstklassifikation, chatbots, talegenkendelse, ordbøger, ontologier, metadataklassifikation, søgemaskiner, maskinoversættelse, værktøj til tekstanalyse, frontends til komplekse lingvistiske backends.
3, 4, 6	Typer af sproglige data	*Fagsprog:* Ja: 5 - Nej: 6 *Alment sprog:* Ja: 10 - Nej: 3 *Nationalsprog:* Dansk: 13 - Andre nordiske: 6 - Engelsk: 5 - Andre: 3
8	Udviklingsmiljø	*OS:* Windows: 4 Linux: 4 Windows og Linux: 3 MacOS og Linux: 1 *Programmeringssprog:* Python, Visual Studio, C#, PHP, Javascript, Java, Perl, Prolog m.fl. *Softwarebiblioteker, IDE'er etc.:* NLTK, gensim, spaCy, scikit, pandas, Visual Studio, apache, solr, Netbeans, Tensorflow, DyNet Tesseract, Tika, BlackLight, Apertium, m.fl.

9-15	Valg af sprogværktøjer	*Danske resurser:* iLex, DanNet, CST's værktøjer *Almene resurser:* espeak iLex, udtaleleksikon, hunspell, brilltagging *Kriterier for valg af nyt software (herunder open-source):* Kontinuitet (eksisterende markeds krav): 5 Udelukkende efter funktionelle krav: 2 Open source-kriterium: 1 Flere af disse: 2 Andet: 1 *Hvordan træffes valg af software?* Kontinuitet (skal understøtte eksisterende marked): 5 Udelukkende efter funktionelle krav: 2 Open source-kriterium: 1 Flere af disse: 2 Andet: 1 *Ville du træffe de samme valg i dag?* Ja: 5 - Betinget ja: 3 - Nej: 1
16-17	Algoritmiske metoder	Hvilke metoder bruger du? Statistiske metoder og neurale net: 1 Regelbaseret: 5 Begge disse: 4 *Ville du træffe de samme valg i dag?* Ja: 8 - Betinget ja: 1 - Nej: 1
18	Hvad oplever du som den største teknologiske begrænsning?	Typiske svar: "Utilstrækkelige data på dansk", "Brug af komponenter der ikke er udviklet med dansk talegenkendelse som målgruppe", "Talesynteser lyder ikke naturligt", "Udvalget af danske talesynteser er begrænset", "Alt for få medarbejdere pga. dårlige økonomiske rammebetingelser", "Manglende support for eksisterende Java applets","Mangel på domænetagget råtekst", "Svært at finde og udnytte forskning"
19-22	Dansksproglige data	Hvilke dansksproglige data arbejder du med? Tilstandsrapporter, Wikipedia, råt tekstmateriale, rå taleoptagelser, ordbogsdata, ontologidata, annoteret tekstmateriale, morfologiske og semantiske leksika, grammatikker, NSTs efterladte data, in-house resurser, ordlister, parallelsproglige data. *Hvilke dansksproglige data mangler du?* Annoteret talesprog, fagspecifikke tekstkorpora, store mængder råtekst, tekst rettet mod børn og andre grupper, forskellige typer danske ordlister, opmærkede danske korpora, korpora til specifikke sproglige udfordringer som stød og sammensatte ord *Hvilke danske data mangler mest?* Annoteret talesprog på dansk, sprogspecifikke løsninger (fx dansk udtale og morfologi), infrastruktur med kontinuitetsgaranti, talesynteser til forskellige platforme
23-24	Deling af resurser	*Er det en fordel at dele resurser med andre firmaer?* Ja: 8 - Nej: 1 - Ved ikke: 1 *Er det en fordel at dele resurser på open source-basis?* Ja: 6 - Nej:1 - Ved ikke: 3

Spørgeskema for forskere og formidlere

Spørgsmålslisten

Ifølge instruktionen skulle spørgsmål 1 og 2 besvares af alle, 3-14 af forskere, 15-27 af undervisere, 28-32 af formidlere og 33-40 af strategiske beslutningstagere. Deltagerne måtte gerne vælge mere end én rolle. Deltagerne blev lovet anonymitet.

1. Hvilket firma/institution repræsenterer du?
2. Hvad er dit fagområde?
3. Hvad er dine centrale forskningsområder?
4. Forsker du i sprogteknologi (fx metoder, resurser, anvendelser)?
5. Bruger du danske sprogdata i din forskning?
6. Forsker du for private og/eller offentlige organisationer?
7. Hvilke samarbejdsrelationer har du i øvrigt (hvor du deltager som ekspert i dansk)?
8. Hvilke NLP-projekter er du involveret i (hvor dansk spiller en rolle)?
9. Hvilke danske sprogteknologiske resurser har du bidraget til?
10. I dit arbejde, trækker du mest på transfer af resurser/teknologier fra andre sprog, eller arbejder du mest med rent dansk udviklede resurser/teknologi?
11. Kandidater med dybere dansk-lingvistiske kompetencer er efterspurgt på arbejdsmarkedet (data science er ikke nødvendigvis nok) - hvordan forestiller du dig det kan imødekommes fra uddannelsessiden?
12. Hvilke forventninger har du til udviklingen på de sprogteknologiske områder i Danmark de næste 2-5 år?
13. Hvilke udviklinger i det danske samfund ønsker du på de sprogteknologiske områder de næste 5-10 år?
14. Skriv referencer til dine seneste artikler om sprogteknologi
15. Underviser du i emner der inddrager dansk sprogteknologi?
16. Hvilke typer af sprogteknologi berører din undervisning?
17. Hvilke modtagere har du?
18. Kommer du ind på dansk sprog og sprogdata i din undervisning?
19. Beskriv de ph.d.-projekter du kender til, som inddrager sprogteknologi (dansk og andre sprog)
20. Er der i din organisation (fx universitetet) kurser med fokus på sprogteknologi?
21. Er der konkrete planer om at lave nye kurser?
22. Hvilke samarbejdsrelationer har du med hensyn til undervisning (hvor du deltager som ekspert i dansk)?
23. I dit arbejde, trækker du mest på transfer af resurser/teknologier fra andre sprog, eller arbejder du mest med rent dansk udviklede resurser/teknologi?
24. Kandidater med dybere dansk-lingvistiske kompetencer er efterspurgt på arbejdsmarkedet (data science er ikke nødvendigvis nok) - hvordan kan det imødekommes fra uddannelsessiden?
25. Hvilke elementer af dansk sprogteknologi kan du forestille dig i fremtidige uddannelser?
26. Hvilke forventninger har du til udviklingen på de sprogteknologiske områder i Danmark de næste 2-5 år?
27. Hvilke udviklinger i det danske samfund ønsker du på de sprogteknologiske områder de næste 5-10 år?
28. Hvilke typer af dansk sprogteknologi oplyser du om?
29. Hvilke modtagere har du?

30. Kommer du ind på dansk sprog og sprogdata i din formidling?
31. Kandidater med dybere dansk-lingvistiske kompetencer er efterspurgt på arbejdsmarkedet (data science er ikke nødvendigvis nok) - hvordan forestiller du dig det kan imødekommes fra uddannelsessiden?
32. Hvilke forventninger har du til udviklingen på de sprogteknologiske områder i Danmark de næste 2-5 år?
33. Hvilke typer af dansk sprogteknologi (produkter/anvendelser) handler det om?
34. Hvem er din målgruppe - eller målgrupper?
35. Hvilke formål sigter du mod ved anvendelsen af dansk sprogteknologi?
36. Hvilke sprog arbejder du med professionelt?
37. Hvilke samarbejdsrelationer har du med andre strategiske beslutningstagere (hvor dansk sprogteknologi spiller en rolle)?
38. Kandidater med dybere dansk-lingvistiske kompetencer er efterspurgt på arbejdsmarkedet (data science er ikke nødvendigvis nok) - hvordan forestiller du dig det kan imødekommes fra uddannelsessiden?
39. Hvilke forventninger har du til udviklingen på de sprogteknologiske områder i Danmark de næste 2-5 år?
40. Hvilke udviklinger i det danske samfund ønsker du på de sprogteknologiske områder de næste 5-10 år?

Besvarelser

	Spørgsmål	Resumerede besvarelser
1	Hvilket firma/institution repræsenterer du?	KU, DTU, Nors, Professionshøjskolen Absalon, DeiC, Alexandra Inst., Oticon, DSL, Det Grønlandske Sprognævn, ITU
2-3	Hvad er dit fagområde/ forskningsområde?	Datalingvistik, sprogteknologi, fonetik, leksikografi, korpuslingvistik, ingeniørvidenskab, datalogi & AI, audiologi, datamanagement, offentlig digitalisering
4	Forsker du i sprogteknologi?	ja: *11* - til en vis grad: *3* - nej, men beslægtede områder: *2* - nej: *1*
5	Bruger du danske sprogdata?	ja: *13* - til en vis grad: *1* - nej: *1*
6-7	Er dine samarbejdsrelationer i privat eller offentligt regi?	private: *1* - offentlige: *13* - både private og offentlige: *1*
8-11	Hvilke typer af dansk sprogteknologi er du involveret i?	Tekstklassifikation, resurseportaler, parsing, tagging, opmærkning af sprogdata (tekst, tale, multimodale), talesyntese, sproglæring
12, 26, 39	Hvilke forventninger har du til udviklingen i Danmark de næste 2-5 år?	Generelt forventes fremgang i teknologiske muligheder (forbedret taleteknologi, flersprogligt værktøj, end-to-end-systemer, AI). Om Danmark vil få del i fremgangen, afhænger af statens investering, både på uddannelse og tilgængelighed af sprogresurser.
13, 27, 40	Hvilke udviklinger i det danske samfund ønsker du de næste 5-10 år?	Dansk skal understøttes mindst lige så godt med sprogteknologi som de øvrige sprog i Europa. Der ønskes stabil politisk bevågenhed og varig støtte til offentlig forskning og undervisning.
14	Dine publikationer	Ikke relevant
15-20	Undervisning inden for dansk sprogteknologi?	Kun rudimentære berøringer med danske sprogdata: kurser i almen korpuslingvistik, ontologi, NLP-programmering, sentimentanalyse
21, 22, 23, 31	Er der konkrete planer om nye kurser/projekter inden for dansk sprogteknologi?	Ingen svarer ja

24, 25, 38	Hvordan styrkes uddannelse af sprogteknologer med dansk kompetence?	Oprettelse af nye uddannelser i sprogteknologi og nye kurser i eksisterende uddannelser. Inddragelse af dansk sprogteknologi som redskab i andre uddannelser.
28, 29	Hvilke typer af dansk sprogteknologi oplyser du om?	Taleteknologi, ordforslag, læsbarhed, eye-tracking, tekstanalyse, generering, m.m., rettet mod universitet og internetbrugere.
30-37	Hvilke sprogteknologiske emner berører din formidling?	Generelt henvises til de udfordringer som dansk skrift og tale giver for automatisk sproganalyse.

BILAG 2
Workshop om terminologi

Workshoppen om terminologi fandt sted den 12.12.2018. Der deltog i alt 28 repræsentanter for danske offentlige institutioner og private virksomheder.

Forud for workshoppen var der udsendt et omfattende spørgeskema til ca. 100 virksomheder og offentlige institutioner. Det blev besvaret af 61 personer, dvs. en svarprocent på lidt over 60. Besvarelserne var meget jævnt fordelt med et lille flertal af offentlige institutioner på 56 %. Resten var private virksomheder eller organisationer.

Svarene kom bl.a. fra

- Myndigheder (styrelser, ministerier, Folketinget), Regioner, Det Europæiske Regionsudvalg, Danmarks Statistik, Nationalbanken
- Virksomheder (KMD, banker, pensionsvirksomheder, produktionsvirksomheder, konsulentvirksomheder, it-udviklere, oversætterbureauer)
- Højere læreanstalter (inst./centre vedr. sprog og kommunikation)
- A-kasser
- Sprog- og kommunikationskonsulenter
- Tolkeforeninger og -virksomheder, netværk af sprogmedarbejdere
- Europa-Parlamentet, Europa-Kommissionen, Oversættelsescentret for Den Europæiske Unions Organer

Deltagerne arbejdede overvejende med dansk, engelsk og tysk.

Det blev tydeligt at størsteparten arbejder med termsamlinger jf. nedenstående definitioner:

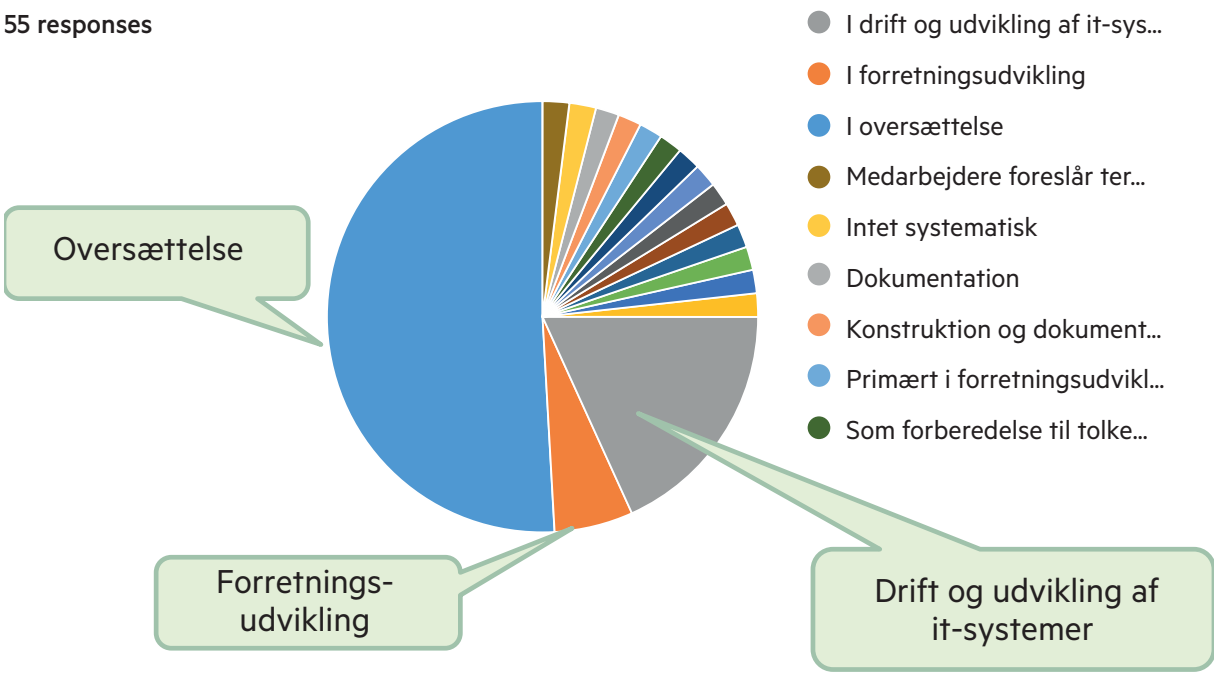
Term	Definition	Note
termsamling	samling af terminologiske informationer	Kan være lagret i og tilgås via fx Excel, Word, PDF og web-grænseflader.
termbase	database som indeholder en termsamling	Fx en virksomhedsintern termbase, som stilles til rådighed for en specificeret mængde af brugere.
terminologihåndteringssystem	software-system som anvendes til adgang til samt lagring og vedligeholdelse af en termsamling	Kan fx være et databasesystem med indbygget grænsefladeprogrammel eller en grænseflade udviklet i et særskilt programmel, som trækker på en termsamling.
termbank	en eller flere termbaser og den institution som administrerer og gør indholdet af disse offentligt tilgængeligt	Stilles til rådighed for en bred og uspecificeret mængde af brugere.

Interne termsamlinger ser ud til at være mest udbredt. Men der findes også en del virksomhedsinterne termbaser. Blandt termbankerne er det især EU's termbank IATE der bruges. Blandt de offentligt tilgængelige termbaser fremhæves den medicinske termbank SNOMED, socialebegreber.dk og sundhedsvæsenets begrebsbase.

En interessant case der understreger behovet for samordning af terminologi på tværs af institutionerne, udgør universiteternes uddannelsesterminologi, hvor KU, AAU, CBS og AU har hver deres egen termbase. Her arbejdes der pt. på at skabe et fælles system, hvilket vil kunne medføre en betydelig resursebesparelse. Behovet for adgang til klare definitioner af begreber forekommer ikke blot blandt sprog- og kommunikationsmedarbejdere og sagsbehandlere, men i lige så høj grad blandt dataarkitekter, ingeniører, forretningsspecialister og it-udviklere.

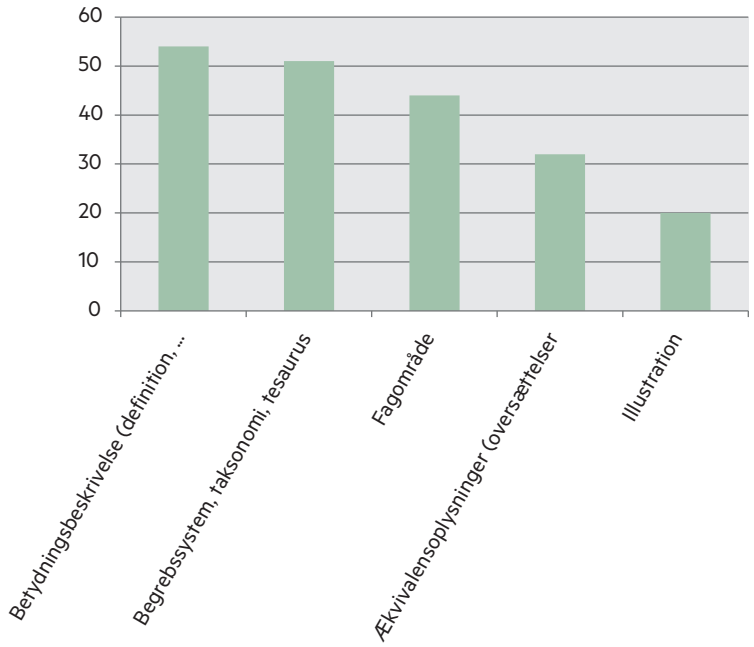
Oversættelse udgør fortsat det største område hvor der er behov for terminologi. Ca. 50 % af de adspurgte har angivet dette som den mest hyppige sammenhæng hvor der er behov for terminologi- eller begrebsarbejde. Men det er tydeligt at områder som drift- og udvikling af it-systemer og forretningsudvikling vinder frem. De udgør sammenlagt 20 % af besvarelserne.

I hvilken sammenhæng foregår terminologi-/begrebsarbejdet?

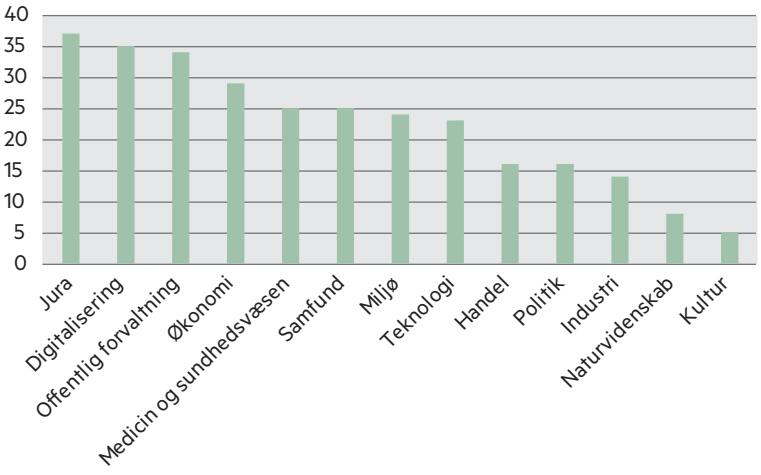


Både af besvarelserne af spørgeskemaet og af tilkendegivelserne i den efterfølgende bearbejdning af svarene på workshoppen fremgår det tydeligt at der er et stort behov for at få samlet terminologi og begreber og deres definitioner et fælles sted i en national termbank.

De oplysningstyper der er størst behov for, er betydningsoplysninger (definitioner), ækvivalente termer (fx forkortelser, varianter og oversættelser til andre sprog) og oplysning om fagområde. Overraskende mange meldte også om behov for en strukturering af begreberne i forhold til hinanden i et begrebssystem, en taksonomi eller en tesaurus. Det er således ikke blot termerne og deres oversættelse der er behov for, men også information om deres indbyrdes sammenhæng.



Blandt de fagområder som hyppigst nævnes som dem der er størst behov for, er jura, digitalisering, offentlig forvaltning, økonomi og sundhed.



Workshoppens konklusioner

Der var bred enighed om at en national termbank vil

- styrke det danske fagsprog og det danske sprog generelt
- bidrage til bedre kommunikation både i den offentlige og i den private sektor
- medvirke til at sikre digitaliseringsklar lovgivning
- skabe bedre effektivitet i offentlige digitale systemer.

Workshoppens vigtigste anbefalinger var:

- Udviklingen af en dansk termbank skal være behovsdreven, dvs. det skal være de fagområder og institutioner der har et konkret behov, der skal gå forrest med hensyn til at beskrive termer og lægge dem i termbanken.

- Alle former for terminologi skal kunne finde plads i termbanken så længe det er tydeligt markeret hvilken kvalitet og hvilken status termerne har.
- Det offentlige kan med fordel være drivkraft for udviklingen, men både den offentlige og den private sektor bør få adgang til data og er velkomne til at bidrage med data.
- Termbanken bør være åben for alle fra offentlige og private virksomheder til uddannelsesinstitutioner på alle niveauer og til den enkelte borger og have passende grænseflader til disse brugergrupper.
- Termbanken skal administreres og kvalitetssikres af fagfolk, og der skal etableres et frugtbart samarbejde med de faglige ildsjæle rundt omkring i institutionerne.
- Der skal skabes gode forbindelser mellem termbanken og den encyklopædiske viden der er beskrevet andre steder, fx i Den Store Danske Encyklopædi, i Trap Danmark, i Sprog- og Litteraturselskabets begrebsordbog og i semantiske beskrivelser som DANnet og FRAMEnet.

BILAG 3
Workshop om automatisk oversættelse

Workshoppen om automatisk oversættelse fandt sted den 8. oktober 2018 i Dansk Sprognævns lokaler i København. Vært for workshoppen var Dansk Sprognævn og den danske EU-repræsentation idet workshoppen foregik i regi af EU's Connecting Europe Facility (CEF) som bl.a. står bag projektet European Language Ressource Coordination (ELRC).

I alt 30 deltagere fra danske offentlige og private institutioner, fra fagforeninger og fra ELRC lyttede til danske og internationale foredrag om automatisk oversættelse og diskuterede hvordan det bedst kan sikres at sprog med få millioner sprogbrugere også kan blive oversat automatisk med en høj kvalitet af EU's sprogservice eTranslation og andre. Fra EU's side understregede man især behovet for at fjerne sprogbarrierer i det digitale indre marked og fx gøre det muligt at drive e-handel på alle EU-sprog og at levere offentlige serviceydelser på nettet til alle EU-borgere på hver deres sprog.

Det blev konkluderet at der i store dele af den offentlige sektor stadig mangler bevidsthed om hvor værdifulde tekster og terminologier er for at man kan skabe bedre oversættelser og bidrage til forskning og udvikling af sprogteknologi for dansk. Mange institutioner indser ikke fordelene ved at oprette og vedligeholde tekst- og terminologidatabaser, eller de kan ikke finde de nødvendige resurser til disse opgaver. Nogle arbejder systematisk med terminologi, men ikke med tekstdata. Mange offentlige institutioner har outsourcet deres oversættelsesopgaver til private leverandører uden at træffe de nødvendige foranstaltninger til at bevare, organisere og genbruge de oversatte data til andre formål.

For at EU's oversættelsessystem og en kontinuerlig indsamling og deling af danske tekstdata skal blive en integreret del af arbejdsgangen i danske offentlige institutioner, bør det også være muligt for private leverandører at anvende EU's oversættelsessystem hvis de oversætter EU-relaterede tekster eller løser andre typer af oversættelsesopgaver for offentlige institutioner.

Et offentlig-privat samarbejde om oversættelse og brug af offentlige sprogdata reguleret af fx reglerne for offentlige udbud kunne sandsynligvis forbedre situationen i Danmark betydeligt.

BILAG 4
Oversigt over de 136 virksomheder og institutioner som har deltaget i udvalgets seminarer og workshops eller på anden måde bidraget til udvalgets arbejde

Alexandra Instituttet
Ankiro Aps
Apple Danmark
ATP, Erhvervsservice og Digitalisering
Ballerup Kommune/OS2
B&L Oversættelsesbureau
BEUMER Group A/S
BotXO
Center for Journalistik, SDU
Center for Sprogteknologi, KU
CIMT, Region Hovedstaden
Civilstyrelsen
Copenhagen Business School, Department of Management, Society and Communication
Corti
Dahl advokater
Danmarks Nationalbank
Danmarks Statistik
Dansk Industri
Dansk Sprognævn
Danske Bank
Danterm Technologies
Det Danske Sprog- og Litteraturselskab
Det Kgl. Bibliotek
Det Nationale Center for Fremmedsprog
Dictus ApS
Digitaliseringsstyrelsen
Digitaliseringsstyrelsen, borger.dk
Doolittle Translation
DR
Edlund A/S
Eggeslevmagle skole, Skælskør
Eniro Danmark A/S
Erhvervsstyrelsen
Europa-Kommissionen
Europa-Parlamentet
European Consumer Centre Denmark/Forbruger Europa
EU's oversættertjenester
Fagbevægelsens Interne Uddannelser, FIU
Favrskov Kommune
First Agenda
Fokus Translatørerne
Folketinget
Folketingets Sekretariat
Forbruger Europa
Forbundet for offentligt ansatte, FOA
Forbundet Kommunikation og Sprog
FORVIR
Google
GrammarSoft ApS
GTS-foreningen
Herlev og Gentofte Hospital
IBM Danmark
Infomedia A/S
Institut for Menneskerettigheder
IT -Universitetet
Kamstrup A/S
KMD A/S
Kommunikation og Teknologi
Konkurrence- & Forbrugerstyrelsen
Københavns Erhvervsakademi
Københavns Professionshøjskole
Københavns Universitet, Center for tekstilforskning
Københavns Universitet, Department of Nordic Studies and Linguistics
Københavns Universitet, Department of Nordic Studies and Linguistics, CST
Københavns Universitet, Kommunikation
Landsorganisationen i Danmark (LO)
LEGO
Lytteskrivning
MAN Diesel & Turbo
MAN Energy Solutions
Max Manus A/S
Medema A/S
MedTech Innovation Consortium
Mette Djørup Software Translation
Microsoft A/S
MIRSK
MV Nordic
Nota
Nuance Communications Ltd
Odense Kommune, IT- og Digitalisering
Oqaasileriffik – Grønlandsk Sprognævn
Ordbog over Dansk Tegnsprog
Oticon A/S
Oversættelsesgruppen
Plandent A/S
PriceWaterhouseCoopers (PwC)
RDFined
Region Hovedstaden, Center for It, Medico og Telefoni
Region Hovedstaden, Sundhedsplatformen
Region Sjælland
Region Sjælland, Sundhedsplatformen
Region Sønderjylland
Rigshospitalet

Samfundslitteratur
Sankt Annæ Gymnasium
Semantix A/S
Skatteforvaltningen
Skattestyrelsen
Slots- og Kulturstyrelsen
Sorø Kommune
Statsministeriets departement
Styrelsen for Arbejdsmarked og Rekruttering
Styrelsen for Udvikling og Forenkling
Styrelsen for Videregående Uddannelser
sundhed.dk
Sundhedsdatastyrelsen
Sundhedsplatformen
Syddansk Sundhedsinnovation
Syddansk Universitet
Saabye
Taxon
Tegnsprogsrådet
tekom Danmark
Terminologigruppen Danmark
Termplus ApS
TextWise
Tino Didriksen Consult
Topdanmark
Translators Café
Translatørbureauet
Translatørbureauet Eunike Hansen
Translatørforeningen
Tryg
TV2
Udenrigsministeriet
UNSILO
Vejle Kommune
Vesthimmerland Kommune
Wizkids A/S
WordMaps
World Translation A/S
Ørsted A/S
Aabentoft A/S
AAC Global
Aalborg Universitet
Aarhus Universitet


Ufarlig!
ALLE HAR BRUG FOR DANSK SPROGTEKNOLOGI
SPROGTEK2018.DK
Rådhuspladsen
Hej
你好
?
Hi, I'm calling to book ...
Je m'appelle: My name is ...
DANSK
Ulven er et rovdyr ...
Folketinget.dk