Kronik: Sproget er digitaliseringens sorte guld

26. september 2016

Politiken Sektion 2 (Kultur) Side 5 (Debat)

PETER JUEL HENRICHSEN, SABINE KIRCHMEIER, BOLETTE SANDFORD PEDERSEN, PHILIP DIDERICHSEN

AUTOMATISERING og digitalisering er store temaer i tidens strategiske politiske arbejde. Udviklingen af samarbejdende robotter er så småt begyndt at revolutionere den industrielle produktion — også i Danmark, der har store internationale spillere på området. Samtidig forventer man sig meget af robotter på velfærdsområdet.

Udviklingen går nu hastigt i retning af at robotter vil blive lige så lette at indstille og styre som apps på telefoner og tablets.

Og den næste milepæl er allerede snublende tæt på: Der forskes nu også i håndfri betjening af robotter ved hjælp af stemmen, en teknologi som allerede findes i biler og digitale assistenter som Apples Siri m. fl.

Designet rigtigt kan stemmestyring og forståelse af spontant dansk blive et overordentlig nyttigt supplement til andre betjeningsmuligheder, f. eks. i service- eller løfterobotter i hjemmet, og betjening af samarbejdende industrirobotter kan gøres endnu mere tilgængelig for folk uden lange uddannelser.

Men hvis det skal give mening i en dansk kontekst, skal betjeningssproget være dansk, og brugeren må ikke tvinges ud i absurd fremmedartet udtale eller andre uvante sproglige konstruktioner for at blive forstået af systemerne.

DET SIGES at den fremadstormende danske robotindustri snart kommer til at mangle ingeniører i tusindvis. Der mangler allerede nu 'sproglige ingeniører', dvs. sprogteknologer og datalingvister, til at tage teknologien skridtet videre og udvikle stemmestyring og sprogforståelse specifikt for dansk.

Stemmestyring er et eksempel på det der kaldes sprogteknologi. Sprogteknologi bruges også i stave- og grammatikkontrol, automatisk oversættelse, kundetilfredshedsundersøgelser gennem big data-analyse af store mængder af f. eks. tweets, terminologihåndtering, vidensdeling og -søgning på dansk, sagsbehandling (f. eks. diktering af journaler), automatisk oplæsning for blinde og ordblinde, udtaletræning for flygtninge og andre immigranter, udvikling af avancerede høreapparater, oplevelseskoncepter — og listen vil blive længere i fremtiden med anvendelser ingen havde forudset.

Erfaringen har vist at Danmark er så lille at danske sprogteknologifirmaer har svært ved at trives på rene markedsvilkår.

De har ikke meget plads i budgettet til at investere i nye tiltag. Sprogteknologi kræver forskning, og forskning er dyrt.

Derfor er meget af den sprogteknologi vi kender fra eksempelvis vores mobiltelefon, blevet udviklet af store internationale firmaer der ikke prioriterer relativt små sprog som dansk så højt, og som derfor heller ikke nødvendigvis benytter sig af danske sprogeksperter og datalingvister.

Regeringen har bebudet en dansk sprogstrategi. Når arbejdet med den går i gang, bør man derfor satse på en strategi der også gør det muligt at udvikle dansk sprogteknologi på danske præmisser.

HVORFOR ER sprogteknologi dyr? Det kan bedst forklares ved at dvæle lidt ved hvordan teknologien egentlig virker.

Det handler teknisk set om analyse af sprogligt input. Tag denne lille stump fiktiv storpolitisk analyse: 'Der er kommet flere såkaldte slyngelstater i verden, og selv de stærke og etablerede magter nærmest ikke at forsvare sig længere'.

Ja, du læste rigtigt — eller gjorde du? Måske skulle der lige et par forsøg til. Sætninger som denne udstiller ords flertydighed: De fleste vil umiddelbart læse *magter* som et navneord (en magt, flere magter...), men finder hurtigt ud af at det ikke kan være sådan sætningen er skruet sammen. Først når magter læses som et udsagnsord (jeg magter...), giver sætningen mening.

SPROGLIG ANALYSE handler om at afklare flertydighed ved at sætte de rigtige etiketter på (som navneord, udsagnsord og de andre ordklasser). Det er denne analyse sprogteknologien programmeres til at foretage automatisk.

Et eksempel er talesyntese, dvs. automatisk oplæsning, som blinde og ordblinde i Danmark hver dag har stor glæde af. Skal ordet *bande* have en udtale der rimer på vande, eller en der rimer på vante? Det kan løses ved at sætte ordklasser på. Hvis bande er et udsagnsord, rimer det på vande og betyder at bruge skældsord.

Hvis det er et navneord, rimer det på vante og betyder en gruppe kriminelle.

Et andet eksempel kunne være automatisk oversættelse af et ord som *skade*.

Her er det ikke nok med ordklasser. Hvis sprogteknologien skal kunne afgøre om der med skade menes en fugl eller en beskadigelse, er man nødt til at bruge etiketter som 'levende væsen' (skade + 'levende væsen' = fugl) og 'tildragelse' (skade + 'tildragelse' = beskadigelse).

DER ER OVERORDNET set to måder at foretage automatisk sproganalyse på: via regler formuleret ud fra sproglig grundviden og via sandsynlighed beregnet statistisk ud fra andre tekster hvor en grammatikkyndig i hånden har tilføjet oplysninger om ordene. Begge er dyre.

Den statistiske tilgang er den fremherskende i øjeblikket, og forskningen har derfor fokus på indsamling og analyse af store mængder tekst. Store tekstmængder er selve grundlaget for denne type sprogteknologi og dermed en vigtig forudsætning for de internationale it-giganters succes. For når først et system som Google Oversæt eller Apples Siri kører, genererer det en konstant strøm af sproglige data der anvendes i udviklingen af værktøjer til glæde for alle os brugere og ikke mindst it-giganterne selv. Men tekstsamlingerne har offentligheden ikke adgang til.

STATISTISK BASEREDE systemer bliver hurtigt gode til at håndtere hyppigt forekommende ord- og sætningsmønstre på basis af hånd-analyserede tekster. Mindre hyppige mønstre håndteres ved hjælp af såkaldt aktiv læring, hvor systemerne selv hjælper til med at udpege sjældne flertydige ord og ordforbindelser. Her må en menneskelig ekspert i hånden analysere flere eksempler hvor konteksten gør ordene entydige (f. eks. jeg så en skade flyve forbi, jeg fik en skade på min bil osv.).

Disse eksempler fodres systemet så med.

Materialet skal altså i sidste ende analyseres i hånden. Det er den menneskelige indsats der koster penge — både de sprogeksperter der opmærker materialet, og de datalingvister der programmerer de sprogteknologiske værktøjer.

Den anden tilgang til automatisk sproganalyse, den regelbaserede, er også meget dyr da leksikografer og lingvister her skal formulere og programmere en stor mængde grammatiske regler for sproget i et givet anvendelsesområde.

Den gode nyhed er at de forskellige sprogteknologiske værktøjer i vid udstrækning bruger de samme basiskomponenter.

I udviklingen af sprogteknologi er der derfor rig mulighed for at slå flere fluer med ét smæk ved at tænke strategisk og udvikle sprogteknologiske basiskomponenter der kan håndtere dansk. Det ligger der et stort potentiale i.

DEN OFFENTLIGE sektor er en af de største — og i fremtiden mest afgørende — brugere af danskudviklede it-systemer, ikke mindst sprogteknologi. Kommuner og regioner har tilsammen et enormt økonomisk volumen og står for næsten halvdelen af det danske bruttonationalprodukt.

Den enkelte kommune er imidlertid, økonomisk set, en svag agent, som sjældent tør stå alene med udvikling af software og it-værktøj. Derfor har over halvdelen af de danske kommuner sluttet sig sammen i en interessekreds kaldet Det Offentlige Digitaliseringsfællesskab, ofte omtalt som OS2. Blandt OS2's undergrupper (dem er der mange af) er OS2Talk særligt interessant i et sprogteknologisk perspektiv.

Gruppen består blandt andet af de kommuner der har praktisk erfaring med automatisk talegenkendelse.

Særligt de danske sprogkomponenter har været problematiske i de talegenkendere kommunerne har kunnet vælge imellem. De er alle udviklet af det amerikanske firma Nuance Communications, og når de ligger hos en enkelt producent, opstår der hurtigt monopollignende tilstande, og det er vanskeligt at kontrollere sprogkomponenternes kvalitet.

OS2Talk blev derfor oprettet med det formål at undersøge hvordan et alternativ kan skabes i et bredt samarbejde mellem de mange interessenter.

AT LAVE GOD talegenkendelse som kan nedskrive danske talte sætninger korrekt, kræver ca. 500 timers udskrevet og analyseret tale samt en sprogmodel (en statistisk tabel over samforekomst af ord) baseret på en tekstmængde på mindst 500 millioner ord der afspejler de fagområder som den færdige talegenkender skal fungere i.

Den slags sprogkomponenter udgør en engangsinvestering da de kan genbruges når først de er udviklet.

På lignende vis udarbejder man maskinoversættelsessystemer ved hjælp af store mængder oversat tekst.

Hvis den slags sprogkomponenter var offentligt tilgængelige, ville kommunerne være i en helt anden situation. Der ville være flere produkter på markedet fra danske producenter, og de udenlandske firmaer ville formentlig være mere interesseret i at integrere de danske komponenter i deres produkter. Dermed ville kvaliteten, og dermed også produktiviteten, ligge på et langt højere niveau end før.

Hvis vi kan få denne solidariske tilgang op i national skala, vil der være lagt et godt fundament for brugen af dansk i mange nye digitale sammenhænge med både kulturelle og økonomiske gevinster til følge.

OG DET KAN lade sig gøre. I Nederlandene har man igennem flere år arbejdet med en strategi for udvikling af en åben, dvs. offentligt tilgængelig, værktøjskasse med basale sprogteknologiske komponenter og dedikerede programmer.

I dag optræder nederlandsk typisk på niveau med større sprog som fransk, tysk og spansk med hensyn til sprogteknologisk dækning.

Og senest har Letland (med sine kun 2 millioner lettisktalende) markeret sig med en flot sprogteknologisk satsning i forbindelse med Letlands EU-formandskab i 2015. Det resulterede bl. a. i en automatisk flersproget informationskiosk der åbnede landet for internationale journalister på en anderledes ukompliceret måde end tidligere. Den udviklede maskinoversættelse (der i øvrigt var selveste Google Oversæt overlegen) blev samtidig brugt til at åbne landet indadtil idet alle offentlige hjemmesider blev forsynet med automatisk oversættelse til russisk til gavn for den relativt store russisktalende befolkning.

I begge eksempler er de store fremskridt blevet mulige gennem strategiske satsninger og vilje til offentlig investering.

DET BEDSTE man kan gøre for at styrke dansk sprogteknologi, er at sørge for at de mest basale komponenter til de forskellige sprogteknologiske værktøjer bliver udviklet, gjort tilgængelige og vedligeholdt løbende som en del af den offentlige infrastruktur.

Tænk på den måde geodata og oplysninger om flyafgange, togtider mv. i dag stilles gratis til rådighed for it-udviklere der så kan levere brugbare mobiltjenester og andre former for service til borgerne.

Herfra vil små som store, offentlige som private virksomheder kunne bruge deres ressourcer på kreativ (videre) udvikling og raffinering af sprogteknologien frem for at bruge tid og kræfter på basiskomponenterne.

Det vil give en højere kvalitet og en hurtigere udvikling og skabe en underskov af firmaer med direkte ekspertise og interesse i dansk — en dansk sprogteknologisk industri som får mulighed for at konkurrere med de internationale it-giganter.

LIGESOM DEN danske vindmølleindustri havde brug for en strategisk offentlig indsats for at opnå sin nuværende stjernestatus på verdensmarkedet, skal der offentlige forsknings-, udviklings- og vedligeholdelsesbevillinger til for at stimulere udviklingen af en dansk sprogteknologisk industri. Her har sprogteknologien i Danmark generelt haft det svært i de senere år. Der har ikke været en strategisk satsning på dette område siden udviklingen af dansk talesyntese i 90'erne, og de bevillinger der er givet siden, er ikke blevet givet efter en samlet strategisk plan, som man har set det i Nederlandene.

Sproget, herunder både almindeligt hverdagsdansk og i særdeleshed al den specialiserede fagterminologi som dansk også rummer, er sprogteknologiens råstof — og en stor, men 'mørk' ressource for big data-feltet.

Mange taler om big data-analyse, men endnu er de færreste klar over potentialet i automatisk at trække information ud af store tekstmængder. Og sproget er en usædvanlig ressource, for jo mere vi bruger det, des mere bliver der af det. Jo mere sprog teknologien har til rådighed, des bedre kan den blive — i hænderne på kompetente specialister, forstås.

TEKST OG TALE skal indsamles, formateres og opmærkes for at kunne indgå i udviklingen af sprogteknologi til gavn for os alle sammen. Og det skal gøres løbende, for sproget ændrer sig løbende. For at sikre den nødvendige kontinuitet er der brug for en åben national sprogbank med tekstdata og basiskomponenter som man allerede har det i Norge, Sverige og Finland.

En strategisk sprogteknologisk satsning bør selvfølgelig også omfatte uddannelse, bl. a. af forskere der kan forske i og udvikle teknologien. Selv om der i dag forskes i sprogteknologi på et højt internationalt niveau i Danmark, uddannes der meget få sprogteknologer og datalingvister.

Der findes ikke en egentlig uddannelse i sprogteknologi, kun valgfag og delmoduler ved enkelte uddannelser som kandidatuddannelsen it & cognition ved Københavns Universitet. Men vi har brug for flere eksperter der dels er gode til computerprogrammering, dels har en god indsigt i dansk sprog, på både bachelor- og kandidatniveau.

SPROGTEKNOLOGI har for nylig været debatteret af kulturordførerne i Folketinget, og kulturministeren har bebudet en sprogteknologisk høring til efteråret.

Men feltet hører rettelig hjemme i en større dagsorden om digitalisering og automatisering, så der er behov for endnu bredere bevågenhed, ikke mindst fra forsknings- og uddannelsespolitisk side.

Fra Uddannelses- og Forskningsministeriet er der flere gange blevet bebudet en samlet national sprogstrategi. Den bør indeholde en strategi for hvordan man udvikler, vedligeholder og anvender værktøjer som støtter, at dansk bliver brugt smidigt på alle de områder hvor det er uundværligt — ikke mindst den offentlige sektors mange servicefunktioner rettet mod den almindelige borger.

Hvis dansk skal forblive en varig og voksende ressource også i det frembrydende digitale Danmark, bør vi ikke overse det i de næste års strategiske prioriteringer.