12.9.2012

Fagre, aktuelle verden – talegenkendelse i Folketinget

Anne Jensen, redaktionssekretær

Referatet af Folketingets forhandlinger er siden efteråret 2007 blevet fremstillet ved hjælp af talegenkendelse, dvs. et system, der omsætter tale til tekst. Hvorfor valgte Folketingets Administration at indføre talegenkendelse, og hvordan fungerer tale til tekst-systemet?

Fra stenografi til talegenkendelse

Siden indførelsen af grundloven i 1849 er talerne i Folketinget (og tidligere også i Landstinget) blevet refereret i deres fulde længde. Frem til midten af det 20. århundrede sørgede stenografer for at fastholde politikernes ord og sætninger. Da det blev svært at rekruttere stenografer, der kunne arbejde i det høje tempo og med den kvalitet, der krævedes, gik man i 1968 over til at optage forhandlingerne i Folketingssalen på spolebånd, som blev skrevet ud på elektrisk skrivemaskine – fra 1985 på computer – og derefter redigeret.

I 2000 blev spolebåndoptagelserne erstattet af et digitalt optagelsessystem, og frem til 2007 blev referatet produceret i to tempi: En medarbejdergruppe, tekstbehandlerne, skrev ordret af, hvad der blev sagt på Folketingets talerstol; desuden opdelte de teksten i afsnit og satte komma, punktum m.v. En anden medarbejdergruppe, referenterne, redigerede talerne; det vil sige, at de bl.a. omsatte talesproget til talesprogsnært skriftsprog.

I starten af 2000-tallet skulle man udskifte det daværende it-system til indskrivning og redigering, fordi det var baseret på en forældet teknik. I den forbindelse valgte Folketingets Administration i 2003 at indhente viden om tale til tekst-systemer anvendt i andre parlamenter, der er sammenlignelige med det danske. Baggrunden var bl.a., at det blev stadig sværere at rekruttere og fastholde kvalificerede tekstbehandlere under de gældende ansættelsesbetingelser.

Dengang var der endnu ikke udviklet et dansk tale til tekst-system, og man kontaktede derfor Videnskabsministeriet, som i samarbejde med Kulturministeriet, DR og TV 2 stod for udviklingen af et dansk system. Desuden fik man Center for Sprogteknologi, Københavns Universitet, til at foretage en afdækning af markedet. Konklusionen var, at det ville være muligt at få leveret og implementeret et dansk tale til tekst-system af tilfredsstillende kvalitet inden udgangen af 2007.

På den baggrund besluttede Folketingets Præsidium i 2005, at referatet af forhandlingerne for fremtiden skulle produceres ved hjælp af et tale til tekst-system, der skulle udvikles til Folketinget. Systemet blev taget i brug ved fremlæggelsen af finanslovforslaget i efteråret 2007.

Udviklingen af Folketingets system

Vores system er udviklet og implementeret af Sirius IT i samarbejde med Prolog Development Center A/S og Max Manus A/S. De to sidstnævnte firmaer har specialiseret sig inden for bl.a. talegenkendelse og indgår i partnerskab med Philips Speech Recognition Systems, der har leveret software til systemet.

I opstillingen af krav til systemet, herunder brugervenlighed, kunne vi trække på de erfaringer, som Vejle Sygehus, DR og TV 2 havde gjort sig med udviklingen af de tale til tekst-systemer, de tidligere havde indført som de første i Danmark. Derudover er brugerne af systemet blevet inddraget i udviklingen af vores system for at sikre, at brugergrænseflader o.a. på den bedst opnåelige måde imødekommer brugernes behov.

Systemet er integreret med Folketingets digitale lydoptagelsessystem, Folketingets dokumentbehandlingssystem, Folketingets hjemmeside, ordbøger og andre arbejdsredskaber.

Tale til tekst-systemets opbygning

Folketingets talegenkendelsessystem består af tre komponenter:

- en akustisk profil af referenten
- en ordbog
- en sprogmodel.

Den akustiske profil er en statistisk model/matematisk repræsentation af sprogets lyde. Systemet er leveret med en standardmodel for dansk, som tilpasses den enkelte referent, ved at han/hun indtaler 1 times tekst, inden systemet tages i brug. Referentens udtale af fonemerne i dansk gemmes i den statistiske model, og derved bliver den akustiske profil specifik for den enkelte referent. Den forbedres, i takt med at referenten bruger systemet.

Ordbogen indeholder ca. 110.000 ord og er opbygget på baggrund af de sidste 10 års referater; den er fælles for alle brugere. Vi opdaterer løbende ordbogen med nye ord, f.eks. *rødvinsreform* og *CO_2-udslip* og navne på nye folketingsmedlemmer. Hvert ord er opført i skrevet form og i en fonetisk transskription (lydskrift), og alle nødvendige former af et ord skal lægges ind i ordbogen, hvilket f.eks. betyder, at vi både opfører *rødvinsreform, rødvinsreformen* og *rødvinsreformens*.

Sprogmodellen er en statistisk baseret beskrivelse af, hvordan ord kan følge efter hinanden, dvs. sandsynligheden for, at et bestemt ord følger efter et andet. Den er udviklet ved at træne systemet på de sekvenser af ord, som forekommer i de sidste 10 års folketingsreferater. Der er således ingen syntaks i sprogmodellen, og det betyder, at der ikke indgår nogen som helst information om, hvilken ordklasse et ord tilhører, og hvilken syntaktisk funktion et ord har i en given streng af ord. Det betyder, at systemet ikke skelner mellem, at ordet *lærer* er et verbum bøjet i nutid og udsagnsled i sætningen *De unge lærer ikke nok matematik*, og at ordet *lærer* er et navneord og indgår i grundleddet i sætningen *En ung lærer har lang forberedelsestid*.

Sprogmodellen forbedres, efterhånden som man anvender systemet, dvs. systemet lærer af at behandle nye strenge af ord.

Hvordan omsætter systemet tale til tekst?

5 minutter efter at en tale er blevet holdt på Folketingets talerstol, kan referenten lytte til den og begynde indtalings- og redigeringsarbejdet. Referenten lytter til en eller flere sætninger og indtaler derefter sætningen, samtidig med at han/hun redigerer den.

Den dikterede tale omsættes til tekst, ved at systemet sammenligner strengen af lyde i de dikterede ord med lydene i den akustiske model. Systemet finder den streng, der bedst svarer til den dikterede lyd, og sammenligner så med de fonetiske transskriptioner af ord, der ligger i systemets ordbog, for at finde ord, der bedst svarer til de dikterede. Derefter bruges sprogmodellen til at finde den mest sandsynlige tekststreng. Gennem de processer finder systemet altså den mest sandsynlige overensstemmelse mellem det dikterede og det, der ligger i ordbogen og i sprogmodellen; og systemet omsætter så det indtalte til en tekst på referentens skærmbillede.

Hvis udtalen af det dikterede ord er enslydende med et andet ords udtale som f.eks. *hver* og *vejr*, finder systemet flere ord i ordbogen, hvis transskription svarer til det dikterede ords. Ordene forekommer måske med forskellig frekvens, så systemet bruger derfor også sprogmodellen til at finde den mest sandsynlige tekststreng, som de fundne mulige ord kan indgå i. Systemet afstemmer så de sandsynligheder med hinanden, finder den mest sandsynlige sekvens af ord og omsætter talen til tekst.

Ukendte ord

Det forekommer af og til, at medlemmerne af Folketinget i deres taler bruger ord, som ikke ligger i systemets ordbog. Dem kan systemet naturligvis ikke genkende. I stedet omsætter systemet ordet til det ord i ordbogen, hvis udtale ligger tæt på det indtalte ords. Et par eksempler på det er følgende: Referenten indtalte … *kommunikere via Facebook*, systemet omsatte det til … *kommunikere via fisk*. Referenten indtalte *Undtagelse fra reglen gælder tvangskonvertering*, mens systemet genererede teksten *Undtagelse fra reglen gælder Tange Sø og Beijing*. Hverken *Facebook* eller *tvangskonvertering* lå på det tidspunkt i systemets ordbog, så referenten måtte taste de rigtige ord ind i teksten.

Sammensatte ord, der ikke ligger i ordbogen, men hvis led ligger i ordbogen, omsættes til to ord, f.eks. blev ordet *porteføljeaktier* omsat til sekvensen *portefølje aktier*.

Systemet lærer

Som omtalt ovenfor sker der en forbedring af både den akustiske profil og sprogmodellen, i takt med at systemet anvendes. Forbedringen foregår på den måde, at systemet sammenligner den genkendte tekst og

den eventuelt rettede tekst, dvs. rettelser, der er sket ved indtaling. Nye sekvenser af kendte ord vil derefter indgå i den statistiske sprogmodel.

Ord, som systemet ikke har genkendt, fordi de ikke ligger i dets ordbog, og som referenterne derfor har rettet ved indtastning, opføres automatisk på en liste og kan derefter lægges ind i ordbogen.

Problemer ved talegenkendelsen

På trods af at tale til tekst-systemet fortløbende lærer nye sekvenser af ord, er der nogle problemer med genkendelsen. Det gælder især på følgende områder:

1. Homofoner, dvs. ord, der udtales ens, men som har forskellig betydning.

Referenten dikterer f.eks. *Det har hverken fremmet samarbejdet eller de resultater, man har opnået…Er den tanke ministeren fremmed?* Den korte tillægsform *fremmet* og tillægsordet *fremmed* udtales på samme måde, og systemet genererer teksten *Det har hverken fremmed samarbejdet eller de resultater, man har opnået…Er den tanke ministeren fremmed?*

Ved de udsagnsord, hvis rod ender på vokal efterfulgt af *-r*, f.eks. *køre*, udtales navneformen *køre* og nutidsformen *kører* på samme måde. Nogle gange danner systemet den forkerte form som i *Vi kan ikke se, hvordan det skal kunne kører godt* og *Det gælder om, at alle lære at følge reglerne*.

De to forholdsord *af* og *ad* udtales begge *a*, og systemet omsætter af og til dikteret *ad* til *af* som i *Det bringer os ingen vegne at gå videre af den vej*. Dette problem kan til en vis grad undgås, ved at man udtaler *ad* med såkaldt blødt *d*. Men referenterne skulle gerne kunne bruge systemet uden at lægge deres udtale radikalt om.

Også infinitivpartiklen *at* og bindeordet *og* kan udtales ens, nemlig *å*, så vi kan opleve, at systemet danner teksten *Og komme videre i debatten synes ikke muligt nu*, selv om der blev dikteret *At komme videre i debatten synes ikke muligt nu*. Problemet kunne imødegås ved altid at bruge udtalen *åw* af ordet *og*; også det ville gøre indtalingen mindre ubesværet, for den udtale af *og* ligger ikke lige på tungen hos de fleste sprogbrugere.

Systemet skelner ikke altid mellem næsten-homofoner som f.eks. *ven* og *vend*, hvis udtale kun adskiller sig ved, at *vend* udtales med stød og *ven* uden. Det skyldes, at stød ikke indgår i systemets akustiske model, og transskriptionen af ordene i ordbogen omfatter derfor ikke angivelse af stød.

2. Sekvenser af enstavelsesord, hvoraf flere er ubetonede.

Af og til bliver ikke alle enstavelsesord i en sekvens omsat til tekst. Der dikteres f.eks. *Vi har jo lige haft en lignende oplevelse*, som genereres til *Vi lige haft en lignende oplevelse*. Mest problematisk er det, hvis ordet *ikke* mangler i teksten: *Sundhedsministeren har ikke haft området i lang tid* omsættes til teksten *Sundhedsministeren har haft området i lang tid*.

3. Ord, der efterfølges af ord med initialt *s-*.

Referenten dikterer f.eks. *Hr. Peter Christensen sagde det så tydeligt*, mens systemet omsætter det til *Hr. Peter Christensens sagde det så tydeligt.* Eller der bliver indtalt *Ministerens svar lader meget tilbage at ønske*, som omsættes til teksten *Ministeren svar lader meget tilbage at ønske.*

De ovennævnte problemer med genkendelsen eksisterer stadig, 2 år efter systemet blev taget i brug, og altså på trods af at systemets sprogmodel har haft 2 år til at lære de omtalte ordsekvenser at kende.

Fejlene i den tekst, systemet genererer, skal selvfølgelig rettes, inden referatet sættes på Folketingets hjemmeside. Dog er netop de ovennævnte fejltyper lette at læse henover, og derfor kræver arbejdet med tale til tekst-systemet en større grad af koncentration og omhu af referenterne end tidligere.

Helt ny arbejdsmetode
I Folketingstidende findes der ikke mere tekstbehandlere, men udelukkende referenter. Før vi indførte talegenkendelsessystemet, redigerede referenterne ved at læse den udskrift af den rå tekst, som tekstbehandlerne havde fremstillet. Nu skal referenterne først lytte til sekvenser af en tale, og mens de lytter til en sekvens, skal de samtidig redigere for derefter at indtale det redigerede. Det er dermed en ganske anden og ny arbejdsmetode, som referenterne har måttet lære sig selv gennem at arbejde med systemet og udveksle erfaringer.

Arbejdsmetoden kan nemlig ikke sammenlignes med, hvordan en simultantolk arbejder, for referenten skal ikke blot videregive indholdet af, hvad en taler har sagt fra Folketingets talerstol, nej, referenten skal holde sig så tæt som muligt til de sproglige udtryk, som folketingsmedlemmerne har anvendt, og omsætte talesprog til talesprogsnært skriftsprog.

Desuden skal referenten bryde talestrengen op med punktummer og sætte komma og andre tegn, mens han/hun dikterer. Oftest kan referenten ud over syntaksen bruge intonationen (sætningsrytmen) til at afgøre, hvor en helsætning slutter. Andre gange er det ikke så klart, hvor taleren slutter en helsætning, eller hvordan en ledsætning hænger sammen med et efterfølgende sætningsfragment, og referenten skal så overveje, hvordan der kan skabes syntaktisk sammenhæng i talen. Derudover skal referenten rette faktuelle fejl og fortalelser i talerne.

Hvorfor indtaler politikerne ikke direkte i tale til tekst-systemet?
Der er flere grunde til, at systemet ikke kan anvendes i Folketingssalen, så folketingsmedlemmer ville kunne indtale deres taler direkte.

For det første kræver opbygningen af en akustisk profil, at man indtaler mindst 1 times tekst, og en forbedring af profilen forudsætter, at man løbende indtaler i systemet. Nogle folketingsmedlemmer er sjældent på talerstolen, hvilket ville betyde, at systemet ikke ville kunne omsætte deres tale til tekst med et tilfredsstillende resultat.

For det andet begynder opbygningen af den akustiske profil med, at den mikrofon, man indtaler i, indstilles til lydforholdene i det lokale, man arbejder i, og den lydstyrke, man taler med. Hvis lydforholdene og lydstyrken for indtalingen ændres, skal man igen justere mikrofonen. Det vil ikke være muligt i Folketingssalen, hvor lydforholdene ikke kan holdes konstante, f.eks. står talerne ikke i samme afstand fra mikrofonen, og hvor talerne taler med forskellig lydstyrke.

For det tredje opnås den optimale genkendelse, når man taler tydeligt og ikke alt for hurtigt, man skal nærmest tale som en nyhedsoplæser. Det kan man hverken forlange eller forvente at folketingsmedlemmer skal. Eftersom der er taletidsregler under Folketingets forhandlinger, vil en taler udnytte sin taletid så effektivt som muligt, og det kan betyde, at han/hun taler meget hurtigt. Det kan også betyde, at han/hun afbryder sig selv og starter på en ny sætning, laver fortalelser eller kommer til at bruge et forkert ord.

For det fjerde nævner politikerne selvfølgelig ikke, hvor der skal være punktum, komma m.v. i deres taler.

Af disse fire grunde kan systemet ikke anvendes af folketingsmedlemmerne i Folketingssalen. Og selv om de opbyggede en akustisk profil og lydforholdene kunne holdes konstante m.v., skulle deres taler redigeres, herunder rettes for faktuelle fejl og fortalelser, inden de kunne lægges på Folketingets hjemmeside.

Ønsker til et fremtidigt system
Vores tale til tekst-system genkender hurtigt meget tale med bl.a. de undtagelser, der er nævnt ovenfor. For at imødegå de fejl kræves et ganske andet tale til tekst-system, nemlig et system, som er tilpasset det danske lydsystem på den måde, at transskriptionen af ordene i systemets ordbog bl.a. omfatter stød. Desuden burde der indgå information om, hvilken ordklasse ordene i ordbogen tilhører, og hvilken bøjningsform de har. Endvidere skulle den statistisk baserede sprogmodel trænes på et større antal tekster, så den fra starten omfattede flere mulige strenge af ord, og sprogmodellen skulle kombineres med en regelbaseret model, dvs. syntaktisk information om, hvilke sætningstyper der findes i dansk, og hvilke(n) syntaktisk(e) funktion(er) et ord kan have, f.eks. at kun et udsagnsord, der er bøjet i tid, kan være udsagnsled i en sætning.

Et sådant system findes ikke endnu, heller ikke til genkendelse af engelsk; men vi har da lov til at håbe, at forskning i især kombinationen af en statistisk og en syntaktisk baseret sprogmodel, men også i

udvikling af bedre statistisk baserede sprogmodeller vil bringe os nærmere et endnu bedre talegenkendelsessystem i en fager og nær verden.

Denne artikel fremkom første gang i Mål + Mæle for april 2010. Siden har Folketinget besluttet at udskifte Max Manus-delen og Word med en anden editor, der er RTF-baseret, så man ikke længere er afhængig af Microsofts opdateringer af Word og Microsoft Office. Samtidig har Prolog Development Center overtaget hele ansvaret for talegenkendelsessystemet. Selve SpeechMagic-systemet er blevet overtaget af det USA-baserede Nuance, der er førende inden for tale til tekst-genkendelse.