Dansk sprogteknologi

Bente Maegaard

Direktør, cand.scient (matematik/fransk), Center for Sprogteknologi.

Det europæiske Sprogår sætter fokus på Europas mangfoldighed af sprog, også mindre udbredte sprog, på fremmedsprogsindlæring og brug af fremmedsprog, og ved diskussion af disse emner må man nødvendigvis også tage i betragtning at vi lever i netværkssamfundet.

Internettet er i sig selv med til at promovere flersprogethed idet Nettet har information på virkelig mange sprog. Det er almindeligt kendt at selv om der kommer mere og mere information, også på engelsk, på Nettet, så er procentdelen af Internetsider på engelsk for nedadgående. Dette faktum har betydet at f.eks. amerikanere nu er meget mere interesseret i oversættelse fra andre sprog, og er blevet meget mere bevidst om at der findes information på mange sprog der i større eller mindre grad er uden for deres rækkevidde.

Også i Europa, er Internettets flersprogede karakter med til at understrege betydningen af sprogindlæring og af sprogteknologiske værktøjer der kan lette omgangen med sprog.

AltaVista

Hvis man søger oplysninger gennem søgetjenesten AltaVista, kan man få sin tekst oversat ved hjælp af maskinoversættelsessystemet Systran. Der oversættes f.eks. fra engelsk til fransk, tysk, italiensk, spansk, portugisisk. Der er tale om en gratis, meget hurtigt fungerende tjeneste, og man må sige resultatet ikke altid er af meget høj kvalitet. På den anden side er det sommetider forholdsvis godt, og det er helt klart bedre end slet ikke at forstå teksten fordi den er på et sprog man ikke forstår, eller selv at prøve at slå de enkelte ord op.

Se f.eks. nedenstående. Kildeteksten er fransk og kommer fra avisen *Midi Libre*. Der er ikke redigeret i teksten, hverken i kildeteksten eller i råoversættelsen. Man kan bemærke at programmet ikke er i stand til at skelne overskrifter og anden tekst, og at det heller ikke er i stand til at tage hensyn til layout, font etc. Hvis man blot skal forstå teksten, er dette dog af mindre betydning. Et andet problem der falder i øjnene, er at oversættelsesprogrammet har svært ved at genkende egennavne - og lade være at oversætte dem. '*Midi Libre*' bliver til '*Free Midday*', og det giver jo anledning til et billigt grin. Men alt i alt er der tale om en meget acceptabel gengivelse af den franske tekst på et forståeligt engelsk. Vi forstår at det nu endelig sner i Pyrenæerne, og at næsten alle skiområder er åbne.

Pyrénées : ce n'est plus la neige qui manque... mais les skieurs
(Midi Libre)

Après avoir longtemps attendu, les responsables des stations des sports d'hiver pyrénéennes ont enfin vu tomber la neige en quantité. Ce qui a permis d'ouvrir récemment la quasi-totalité du domaine skiable. Le bulletin des Neiges Catalanes (l'association qui fédère les stations d'hiver des Pyrénées-Orientales), publié samedi, indiquait que l'on peut désormais y pratiquer toutes les formes de ski. Pour le ski alpin, le choix est vaste : Espace Cambre d'Aze- Eyne-Saint-Pierre, 24 pistes ouvertes sur 26 ; Cerdagne-Puigmal, 26 / 26 ; Font-Romeu / Pyrénées 2000, 35 / 40 ; Formiguères, 18 / 18 ; Les Angles, 26 / 29 ; Puyvalador 15 / 16 ; Porté-Puymorens, 13 / 17.

The Pyrenees: it is not any more the snow which misses... but the skiers (Free Midday) After having waited a long time, the persons in charge of the stations of the Pyrenean winter sports finally saw falling snow in quantity. What made it possible to open the near total of the skiable field recently. Bulletin of Snows Catalan (association which federates the stations of winter of Pyrenees-Eastern), published Saturday, indicated that one can from now on practise all shapes of ski there. For the Alpine skiing, the choice is vast: Space Cambers of Aze- Eyne-Saint-Pierre, 24 tracks open out of 26; Cerdagne-Puigmal, 26 / 26; Make-Romeu / The Pyrenees 2000, 35 / 40; Formiguères, 18 / 18; Angles, 26 / 29; Puyvalador 15 / 16; Carried-Puymorens, 13 / 17.

Det er naturligvis udmærket at vi kunne oversætte fra fransk til engelsk; men 'Translate'-knappen findes ikke for dansk. Det betyder at hvis udlændinge vil have adgang til tekster skrevet på dansk, så må de selv lære dansk, eller også må de have teksten oversat af et menneske. Der er således ikke er

nær så god adgang for danske tekster til at blive læst internationalt som for f.eks. franske. Sådan har det jo altid været; men det nye er at det ikke *behøver* at være sådan.

Tilsvarende kan danskere som ikke forstår en fremmed tekst, ikke få den oversat til dansk på Nettet, - men dog måske til engelsk hvilket allerede er en hjælp.

I fremtidens netværkssamfund og med Danmark som en førende IT-nation er det en selvfølge at det danske sprog dækkes med f.eks. en automatisk oversættelsesservice til og fra de mest almindelige sprog.

Sprogteknologiske produkter

Sprogteknologiske produkter omfatter alle former for programmer, der hjælper med til at forbedre og effektivisere menneskers arbejde med tekster: stave- og grammatikkontrol, informationssøgning, oversættelse, ordbogsopslag, undervisningsprogrammer, dialogprogrammer, diktereprogrammer, syntetisk tale osv.

Vi har ovenfor fokuseret på den almindelige bruger, men det er også nødvendigt for erhvervslivet og for den offentlige sektor, kort sagt for os allesammen, at have adgang til en bred vifte af sprogteknologiske programmer af god kvalitet, for at lette arbejdet med tekster.

Og det gælder ikke blot tekster, men også tale. Hvis der ikke er gode talegenkendelses- og talesynteseprogrammer for dansk, vil vi selv og den kommende generation skulle nøjes med engelsk tale i underholdningsspil, undervisningsprogrammer mv. Man kan måske sige at det er sundt for børnene at lære noget engelsk tidligt i livet; men det er en meget svag 'begrundelse' for at tillade en sådan majorisering af vores eget sprog. Hvis vi aktivt vil værne om og udvikle vores sprog, så skal vi sikre at dansk kan anvendes alle steder hvor det føles naturligt, også over for børnene. Derfor skal vi se på status for dansk sprogteknologi.

Orddeling

Automatisk orddeling er en integreret del af tekstbehandlingspakkerne; automatisk orddeling foregår enten ved opslag i en stor ordbog, hvor alle ord er lagret med de mulige delepunkter eller ved brug af en algoritme (der kan være regelbaseret eller et neuralt netværk). Ligegyldigt hvilken metode der anvendes, kan man imidlertid ikke regne med 100% korrekte forslag fra programmet. Opslagsmetoden forudsætter at alle ord kan samles i en ordbog; men det er ikke muligt for vi danner hele tiden nye ord. Algoritmemetoden kommer til kort fordi det er meget svært at formulere regler for orddeling som ikke kræver at man forstår teksten. Hvordan skal *vandret* deles? Det kommer vel an på, om der står *planen var vandret* eller *de har vandret i Pyrenæerne*. Selv om man altså ikke får 100% korrekte forslag fra programmet, er der klart tale om en optimerende faktor ved produktion af større mængder tekst der skal deles ved linieskift af hensyn til god udnyttelse af papiret.

Stavekontrol

Stavekontrol er også standard; et stave-

kontrolprogram skal finde alle fejl i teksten, men må helst ikke markere for mange korrekte ord som mulige fejl, og så skal det ved fejl gerne foreslå det ord som man egentlig havde tænkt på. Også her er det vanskeligt, for ikke at sige umuligt, at opnå 100% korrekthed, dels fordi, som ovenfor, ikke alle danske ord kan stå i ordbogen og man derfor må opbygge regler for hvorledes nye ord dannes ved afledning og sammensætning, og dels fordi ord i sig selv kan være stavet rigtigt, men er forkerte i sammenhængen, f.eks. 'det er svært at lærer'. Dansk er her dækket nogenlunde svarende til andre sprog.

Grammatikkontrol

Men når vi kommer til grammatik- og stilkontrol, er det danske sprog ladt i stikken. Der findes grammatikkontrol for engelsk, fransk, tysk osv., men ikke for dansk. Et grammatikkontrolprogram virker ligesom et stavekontrolprogram, blot på næste niveau: programmet undersøger om ordene hænger sammen i sætninger, om de kommer i den rigtige rækkefølge, er bøjet rigtigt (i den engelske grammatikkontrol kan vi danskere f.eks. have glæde af at den undersøger om vi husker s'erne på udsagnsord i 3. person ental (*the first paper concerns*...) mv. En dansk grammatikkontrol ville klare fejl af typen 'lærer'-'lære' som ovenfor.

Hvad er grunden til at vi ikke har grammatikkontrol for dansk når teknologien åbenbart findes og er implementeret for andre sprog? Grunden er at der kun er 5 mill. danskere, plus de udlændinge der gerne vil skrive på vores sprog, og at det åbenbart er for lille et marked til at betale for udviklingen. (For en ordens skyld skal det forholdsvis enkle danske stilprogram *Sprogmagisteren* nævnes; programmet checker ordforrådets stilniveau mv.).

Oversættelse

Der findes nogle få mindre maskinoversættelsesprogrammer, der har dansk som det ene af sprogene (Winger 92, PC Translator, mv.); men man kan f.eks. bemærke at Europa-Kommissionen, der har verdens største oversættelsesafdeling, endnu ikke har noget program der behandler dansk. Kommissionen købte i 1976 maskinoversættelsessystemet Systran og har gennem årene udviklet det til at kunne behandle en række sprogpar, her er dansk det eneste af de (indtil for et år siden) ni EU-arbejdssprog der ikke er repræsenteret. (Nu er vi i godt selskab med svensk og finsk!). Som beskrevet ovenfor er oversættelsesværktøjer utroligt vigtige for de mindre udbredte sprog, det gælder både oversættelse til og fra disse sprog. Oversættelsesværktøjer kan gøre oversættelse af danske tekster til f.eks. engelsk mere overkommelig og dermed bane vejen for at flere danske tanker kan indgå i den internationale debat.

Det skal nævnes, at der findes andre typer af oversættelseshjælp end maskinoversættelse, og her er dansk udmærket dækket. Bl.a. findes mange ordbøger i elektronisk form, så oversætteren har adgang til dem fra tekstbehandlingsprogrammet. Der er også adgang til mange ordbøger og termdatabaser på Internettet. De er ganske vist af svingende kvalitet; men f.eks. Europa-Kommissionens termbase *Eurodicautom*

er et vældigt aktiv. Endvidere findes oversættelseshukommelsesprogrammer i udgaver der kan behandle det danske alfabet. Det er altså kun på det mere avancerede niveau, at vi halter bagud.

For en ordens skyld skal det nævnes at firmaet Lingtech anvender et specialudviklet, avanceret maskinoversættelsesystem ved oversættelse af patentdokumenter fra engelsk til dansk. Dette program er imidlertid ikke generelt anvendeligt, da det er specielt rettet mod patenttekster.

Diktereprogrammer

Diktereprogrammer er systemer der tager talt input til tekstbehandling og gengiver det i skrevet form på skærmen. Teknologien har nået et stade hvor disse programmer faktisk kan bruges. Systemerne kan nu opfatte sammenhængende tale, hvor de første systemer kun kunne genkende isolerede ord. De fleste systemer virker bedst hvis de trænes med en ny brugers stemme og udtalemåde, så man må påregne at skulle bruge lidt tid på at træne programmet inden brug; dette øger kvaliteten betragteligt. Der findes to-tre store producenter af denne teknologi. En af dem, NST - se nedenfor, er gået i gang med at satse på det danske marked, men er løbet ind i økonomiske vanskeligheder, så det er uvist hvornår produktet kommer på markedet.

Talesyntese

Talesynteseprogrammer tager en tekst som input og læser den op. De er nyttige f.eks. for svagtsynede og svage læsere, der kan få læst tekster op i stedet for at skulle læse dem. De er også nyttige for mennesker der er i gang med at bruge hænder og øjne til andre formål end læsning, f.eks. bilkørende der skal have besked om hvad vej de skal køre, eller som gerne vil have læst avisen op på vej til arbejde. Dansk talesyntese er snart fremme på markedet, dels fra Speech-Ware i Ålborg, der har fået støtte til projektet fra IT- og Forskningsministeriet, og dels fra det norske firma Nordisk Språkteknologi, NST, i Voss.

Dialogprogrammer, interaktive talende programmer

I det øjeblik der findes god talesyntese og god talegenkendelse for det danske sprog, kan man udvikle dialogprogrammer til en række formål, herunder informationssøgning, spil, e-handel og undervisning. Det vil forhåbentlig være en realitet inden for ganske få år.

Konklusion

Denne gennemgang af status for dansk sprogteknologi har vist at der er gang i produktionen af sprogteknologi for dansk, men at der stadig er et stykke vej igen. Det er mit håb at Det europæiske Sprogår vil bidrage til at øge opmærksomheden på den sprogteknologi der faktisk findes for dansk, så den vil blive brugt mere; det vil også vil bidrage til at det bliver sat skub i udviklingen af de værktøjer vi stadig savner, - f.eks. en oversættelsesknap på Internettet.