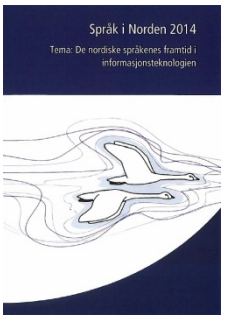
Sprog i Norden

Titel: Sprogteknologi og sproginstitutioner. Hvilken rolle kan sprognævnene spille i forhold til sprogteknologi?

Forfatter: Sabine Kirchmeier-Andersen

Kilde: Sprog i Norden, 2014, s. 133-150

URL: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Søgbarhed
Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Sprogteknologi og sproginstitutioner. Hvilken rolle kan sprognævnene spille i forhold til sprogteknologi?

Sabine Kirchmeier-Andersen

De nordiske sprognævn har igennem flere år arbejdet ihærdigt for at fremme kendskabet til sprogteknologi og for at bringe relevante aktører på området sammen for at styrke det nordiske samarbejde. Sprognævnenes arbejdsgruppe for sprogteknologi i Norden (ASTIN) har bl.a. afholdt flere succesfulde konferencer og workshops med sprogteknologi på programmet. I det følgende vil det blive belyst hvilke resultater dette arbejde har skabt, og hvordan samarbejdet om sprogteknologi kan videreudvikles for at styrke brugen af de nordiske sprog i fremtidens informationsteknologi.

Sprogteknologi, den nordiske sprogdeklaration og nordisk sprogpolitik

Sprogteknologi nævnes ikke direkte i den nordiske sprogdeklaration (Deklaration om nordisk sprogpolitik 2006), men der fremhæves under arbejdsspørgsmål 1 tre applikationer som forudsætter sprogteknologi:

- Internordiske ordbøger i papirform og elektronisk form udarbejdes

- Maskinoversættelsesprogrammer for Nordens samfundsbærende sprog og programmer til søgning i nordiske databaser udvikles (Deklaration om nordisk sprogpolitik 2006, s. 13)

Sprogteknologi er imidlertid ikke begrænset til dette, men er et middel der gør det muligt hurtigere og bedre at opnå flere af de andre målsætninger der opstilles i sprogdeklarationen, og flere de arbejdsspørgsmål som prioriteres, især dem som vedrører parallelsproglighed og mangesprogethed/flersproglighed. Udviklingen af sprogteknologi er også en forudsætning for at man kan nå målet om at gøre Norden til en sproglig foregangsregion.

Sprogteknologiske programmer indgår i undervisningsprogrammer på alle niveauer, i interaktive netsteder, ordbøger, talegrænseflader, stavekontroller osv., og sprogteknologi har længe været blandt de faktorer der kan medvirke til at bevare truede sprog (Crystal 2000, s. 143).

Der er gået 8 år siden sprogdeklarationen blev vedtaget, og netop på det sprogteknologiske område er der sket meget i denne tid. Hvis sprogdeklarationen skulle revideres i de kommende år, ville det være nærliggende at inkludere mere sprogteknologi, heriblandt fx taleteknologi, især talegenkendelse, blandt de applikationer som skulle fremmes for at styrke de nordiske sprog. De mål for sprogteknologi som er opstillet i deklarationen, er kun delvist nået. Således er der kommet enkelte elektroniske ordbøger, fx Islex (http://islex.lexis.hi.is), og der er også udviklet programmer til tværsproglig søgning, som dog kun kører som prototype på Nordisk Ministerråds hjemmeside (http://cst.ku.dk/projekter/projekter_slut/netordbog/). Inden for området maskinoversættelse er der stadig lang vej endnu.

Sprogteknologi spiller også en rolle i de nordiske landes sprogpolitikker. Alle sprogpolitiske redegørelser, hvad enten de er blevet efterfulgt af egentlig sproglovgivning eller ej, indeholder overvejelser om sprogteknologiens betydning. Der er endvidere en generel anerkendelse af at de nordiske sprog som marked betragtet ikke er store nok til at internationale virksomheder som indarbejder sprogteknologi i deres produkter, fx Microsoft (stavekontrol) eller Apple (talegenkendelse), kan motiveres til at investere nævneværdige beløb i produkter af høj kvalitet for disse sprog.

Anbefalingerne i de sprogpolitiske redegørelser går typisk på at opbygge sprogteknologiske infrastrukturer og satse på forskning i sprogteknologiske produkter der tager højde for de specielle karakteristika man finder i de nordiske sprog. Der er med andre ord en klar erkendelse af at sprogteknologi for de enkelte sprog bør styrkes, men der er stor forskel på hvor meget de enkelte lande vælger at investere, fx har Norge og Sverige for flere år siden etableret nationale termbanker, og der arbejdes nu på at oprette permanente sprogbanker til organisering og tilgængeliggørelse af sprogresurser. Språkrådet i Norge har presset på over for de politiske beslutningstagere for at få gennemført planerne om sprogbanker. Etableringen af en permanent sprogbank blev besluttet i 2009 og påbegyndt i 2010. I Sverige er arbejdet ikke så langt fremskredet, men der arbejdes nu på et pilotprojekt om en sprogbank med talesprogsdata. Til

trods for at termbanker og sprogbanker har været en del af de sprogpolitiske anbefalinger i Danmark i de sidste 10 år og forslag om termbanker har været på forskningsministeriets kortlægning af forskningsinfrastruktur og flere gange har været behandlet i Folketinget, er der endnu ikke er blevet givet permanente bevillinger til nogen af delene.

Hvad er sprogteknologi, og hvordan virker det?

Sprogteknologi er en fællesbetegnelse for teknikker der anvendes til at få computere til at bearbejde tekst eller talt sprog. Fra simple programmer til optælling af ord i løbende tekst i 1960'erne har sprogteknologien i takt med udviklingen i informationsteknologien udviklet sig til at omfatte mange forskellige aspekter af sproget og dækker i dag mange forskellige anvendelsesfelter. Sprogteknologi overlapper delvist med andre it-områder som multimedieteknologi og kunstig intelligens. Der er meget forskellige teknikker knyttet til at håndtere input på henholdsvis talt og skrevet sprog, mens de teknikker som bruges til at analysere indholdet er fælles.

Sprogteknologi indgår efterhånden i alle former for informations- og kommunikationsteknologi lige fra undervisningsprogrammer og computerspil til informationssøgning og automatisk oversættelse. Også moderne ordbøger og terminologi- og vidensbaser indeholder sprogteknologiske komponenter. Automatisk oversættelse og tolkning er blandt de største udfordringer for sprogteknologien idet oversættelse kræver en meget avanceret forståelse af både kildesproget og målsproget og af relationen imellem dem.

Figur 1

Mange computerprogrammer, databaser og webapplikationer som involverer sprog, bliver udviklet uden en grundlæggende viden om sprogteknologi, og dette medfører ofte uhensigtsmæssigheder som kunne være undgået hvis producenterne havde involveret mennesker med sprogteknologisk kompetence.

Stort set alle applikationer der involverer sprogteknologi, er skabt ved hjælp af to grundlæggende byggesten: En teknikresurse, typisk et sprogteknologisk program eller softwaremodul, og en sprogresurse, fx store samlinger af talt eller skrevet sprog – det sproglige råstof.

Figur 2

Disse byggesten bygges sammen til applikationer med to grundlæggende metoder, regelbaserede og statistiske. Når der bruges regelbaserede metoder, udarbejder sprogteknologer datamatiske ordbøger, grammatikker og oversættelsesregler eller regler for hvordan tale omsættes til skrift, ud fra deres viden om sprog og ud fra de observationer de kan gøre med udgangspunkt i sprogresurserne. Disse regler bygges så ind i programmer der fx kan analysere og omforme et spørgsmål fra en kunde til en forespørgsel i en database eller generere et svar på et spørgsmål man har stillet til en chatrobot, en service på nettet hvor man kan få besvaret sine spørgsmål af computeren, fx IKEA's chatrobot Anna.

Når der bruges statistiske metoder, udleder et program reglerne automatisk fra de digitale sprogresurser, fx kan sandsynligheden for at et givet ord på et sprog skal oversættes med et givet ord på et andet sprog, beregnes ud fra de statistiske informationer der kan uddrages af en sprogresurse bestående af en samling tekster på et sprog og deres oversættelse til et andet sprog. De beregnede sandsynligheder og regler indgår derefter i applikationer som fx Google Translate.

Begge metoder har fordele og ulemper. Regelbaserede metoder er langsomme at udvikle fordi de er mere manuelle og kræver mange personers arbejde. Til gengæld er de ofte mere generelt anvendelige. Statistiske metoder er markant hurtigere at udvikle, men kun hvis der allerede foreligger digitale sprogresurser i rigelige mængder, og det er særdeles vanskeligt for sprog som færøsk, grønlandsk og samisk (Langgård 2014, Moshagen 2014). Endvidere er der den ulempe ved statistiske metoder at de kun kan analysere og producere den type sprog som anvendes i sprogresurserne. Træner man et maskinoversættelsessystem på fx debatter fra EU-parlamentet, vil de virke fint for politiske debattekster, men komme til kort når man skal oversætte en brugsanvisning til en husholdningsmaskine.

I de sidste 20 år har systemer som er baseret på statistiske metoder, domineret markedet. Mange store applikationer er blevet udviklet, fx systemer som Google Translate, automatisk talegenkendelse og diktering i bl.a. sygehussektoren, fx Nuance/MacManus, og avanceret informationssøgning, fx IBM's søgemaskine Watson (www-03.ibm.com/innovation/us/watson) der har slået alle rekorder i den amerikanske paratvidenquiz Jeopardy og nu bl.a. bliver anvendt i medicinsk forskning og i finanssektoren.

Man er dog ved at nå til et punkt hvor det er vanskeligt at opnå nævneværdige forbedringer af kvaliteten af de statistiske metoder, og et mere avanceret samspil mellem regelbaserede og statistiske metoder synes at være vejen frem. Tendensen går i retning af at kombinere de to metoder således at de regler som udvikles statistisk, finjusteres ud fra en regelbaseret tilgang, eller at man træner systemerne på sprogresurser hvor der allerede er foretaget en regelbaseret analyse, fx tekster som er manuelt opmærket med syntaktiske funktioner eller dependensoplysninger som en slags guldstandard.

Sprognævnenes engagement i sprogteknologi

Sprognævnene er både brugere og producenter af sprogteknologi og sprogteknologiske resurser. Det giver god mening at sprognævnene følger med i den sprogteknologiske udvikling. For det første fordi det ikke kan udelukkes at teknologien kan påvirke sprogbrugen, og for det andet fordi der bliver udviklet viden og værktøjer der støtter sprognævnene i nogle af deres kerneopgaver. De ordbøger, tekstsamlinger og svarsamlinger som sprognævnene selv producerer eller er med til at producere,

udgør en vigtig resurse for andre producenter af sprogteknologi, og det vil være fremmende for udviklingen hvis sprognævnene kan stille disse resurser til rådighed for forskere og udviklere. Sidst men ikke mindst ser sprognævnene det som en vigtig opgave at styrke udviklingen af nordisk sprogteknologi som en del af opfølgningen på den nordiske sprogdeklaration.

Netværket for sprognævnene i Norden (NSN) har nedsat en række arbejdsgrupper som har til opgave at følge op på centrale mål i sprogdeklarationen, heriblandt i 2005 Arbejdsgruppen for sprogrøgt og sprogteknologi i Norden (ASTIN).

ASTIN har følgende opgaver:

- at arbejde for at sprogteknologisk forskning og udvikling får adgang til den nødvendige infrastruktur i form af grundlæggende sprog- og teknikresurser for de nordiske sprog

- at arbejde for at vigtige sprogteknologiske applikationer udvikles for de nordiske sprog, og at disse tilpasses grupper med særlige behov også når den kommercielle interesse savnes

- at gøre opmærksom på teknologiens konsekvenser for skriftsprogsnormen og sproganvendelsen i Norden og at arbejde for at nordisk sprogteknologi har en god sproglig kvalitet

- at styrke det nordiske sprogrøgtsarbejde ved hjælp af sprogteknologi

- at fremme samarbejdet mellem sprognævn, forskning og industri i Norden på det sprogteknologiske område

ASTIN består p.t. af

- Torbjørg Breivik, Språkrådet i Norge

- Rickard Domeij, Språkrådet i Sverige

- Per Langgård, Grønlandsk Sprognævn

- Sjur Nørstebø Moshagen, Sametinget
- Sabine Kirchmeier-Andersen, Dansk Sprognævn

ASTIN har siden 2005 først og fremmest arrangeret konferencer og seminarer for sprognævn, forskere og virksomheder:

2005	Språkverktøy for språkene i Norden, 21.–22. april, Pargas, Finland, ca. 50 deltagere
2006	Språkteknologisk infrastruktur for språkene i Norden, 26. oktober, Göteborg, ca. 60 deltagere (samarbejde med SLTC (Swedish Language Technology Conference)
2007	Sprogrøgt, sprogpoltik og sprogteknologi, 29.–30. oktober, København, ca. 60 deltagere
2008	Oversettingsteknologi i Norden, 23.–24. oktober, Oslo, ca. 45 deltagere
2009	Nordiske perspektiver på CLARIN's infrastruktur for sprogresurser, 14. mai, Odense, ca. 30 deltagere (arbejdsseminar på de Nordiske Datalingvistikdage (NoDaLiDA 2009))
2009	Sprogteknologisk terminologi, 12. juni, København, ca. 30 deltagere (arbejdsseminar i tilknytning til Nordterm 2009)
2010	Språkteknologi for økt tilgjengelighet, Linköping, 27.–28. oktober, 45 deltakere (i samarbejde med SLTC 2010)
2011	Visability and availability of LT resources, 10. mai, Riga, ca. 35 deltakere (arbejdsseminar på de Nordiske Datalingvistikdage (NoDaLiDa 2011))
2012	Taleteknologi – hva har vi, og hva vil vi? Lund, 24.–25. oktober, 55 deltakere (konference i tilknytning til SLTC 2012)
2013	Bidraget til programmet for det nordiske sprogmøde som havde sprogteknologi som tema.

Tabel 1

Derudover har ASTIN fungeret som styre- eller rådgivningsgruppe for forskellige projekter, fx sprognævnenes svarbaseprojekt (www.nordsvar.org) og projekter om kortlægning af nordiske sprogresurser, fx den nordiske vismandsrapport om sprogteknologi (Koskenniemi, Lindén & Nordgård 2007). ASTIN's medlemmer er typisk også involveret i natio-

nale sprogteknologiske projekter, typisk infrastrukturprojekter som fx sprogbanker og i de nordiske afdelinger af EU-projekterne CLARIN og META-NET.

Sprognævnenes sprog- og teknikresurser

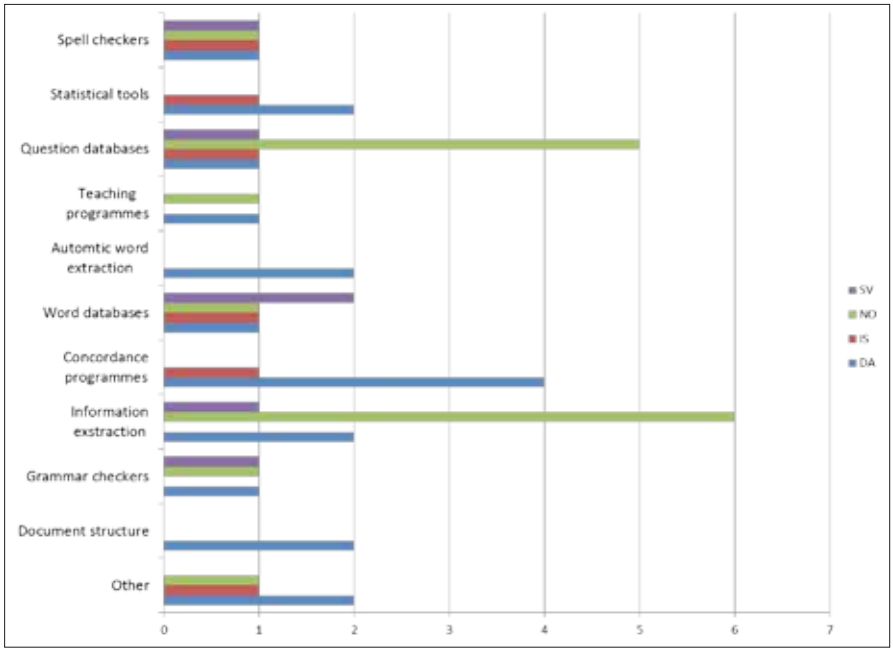
I 2007 foretog ASTIN en undersøgelse af de resurser som de nordiske sprognævn bruger eller selv opbygger, for at få et bedre billede af på hvilke områder der kunne være mulighed for et tættere samspil med forskningsmiljøer og sprogteknologiske virksomheder (Kirchmeier-Andersen 2011).

Undersøgelsen tog udgangspunkt i såkaldte BLARK-koncept (Basic Language Ressource Kit) (Krauwer 1998 og www.blark.org) som opregner de sprogteknologiske resurser der er nødvendige for at sprogteknologi kan udvikles for et givet sprog.

Blandt de teknikresurser som blev kortlagt, var fx parsere (dvs. automatisk morfologisk analyse og sætningsanalyse), programmer til informationsekstraktion, værktøjer til automatisk opmærkning, konkordansprogrammer eller transskription af talt sprog, statistiske værktøjer, stavekontroller, grammatikkontroller, programmer til automatisk ekstraktion af termer og nye ord, orddatabaseprogrammer, programmer til udvikling af sprogøvelser og -tests mv.

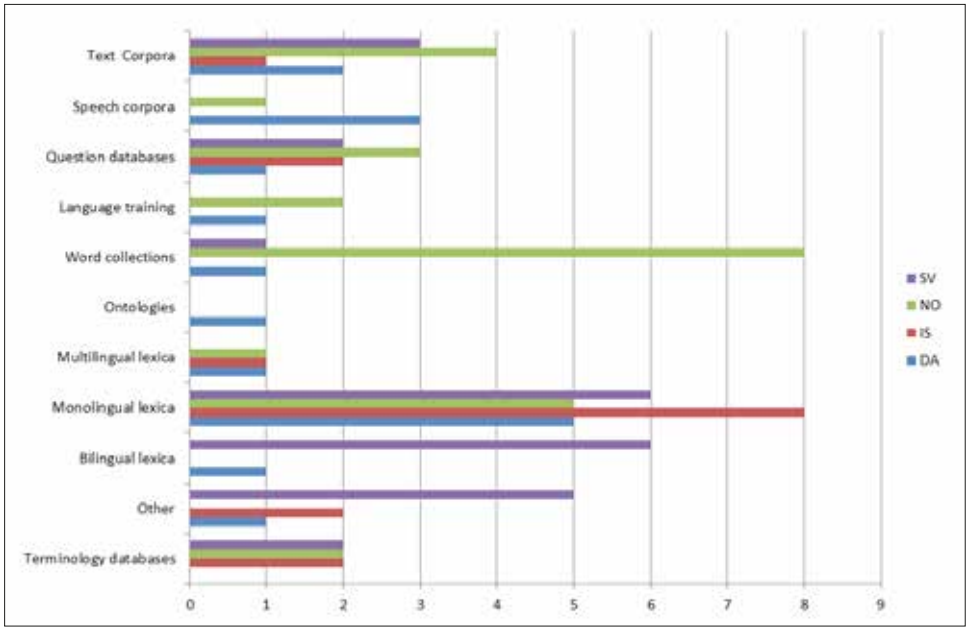
Blandt de sprogresurser der blev kortlagt var fx digitale ordbøger (mono-, bi- og multilingvale), korpusser (tale, tekst, flersproglige, og parallelle), ordnet og tesaurusser, ontologier, svarsamlinger, ordsamlinger, sproglige øvelser, terminologisamlinger. I undersøgelsen deltog Sverige, Norge, Island og Danmark.

Hvad angår teknikresurser, altså sprogteknologisk software, viste det sig at kun stavekontroller, svardatabaser og orddatabaser var i brug i alle fire sprognævn, og at der typisk var forskellige typer af svardatabaser, konkordansprogrammer og programmer til informationsekstraktion til rådighed. Programmer til analyse og generering af dokumentstruktur og automatisk ekstraktion af nye ord var på dette tidspunkt kun i brug i Danmark.



Oversigt over teknikresurser i de nordiske lande

Hvad angår sprogresurser, viste det sig at tekstkorpusser, ordsamlinger og ensprogede elektroniske ordbøger fandtes i alle fire sprognævn, mens fx sprogundervisningsmateriale kun fandtes i Danmark og Norge, mens databaser med terminologiske data kun fandtes i Sverige, Norge og Island.



Oversigt over sprogresurser i de nordiske lande

Dataene blev præsenteret på et seminar i København i 2007 og dannede efterfølgende grundlag for dialog med virksomheder og forskningsmiljøer. Det blev bl.a. tydeligt at sprognævnene ikke selv producerer sprogteknologi i nævneværdigt omfang, men at de gerne indgår et tæt samarbejde med udviklervirksomheder og forskningsinstitutioner om udvikling og afprøvning af nye programmer. Endvidere blev sprognævnene opmærksomme på at de resurser de selv udvikler, kan have stor værdi for forskere og udviklere. En øget resursedeling, fx efter et open source-princip, stiller imidlertid både krav til klarlæggelse af de ophavsmæssige rettigheder til resurserne og til sprognævnenes strukturering af de sproglige data.

Sprogteknologi i de nordiske lande i dag

Selv om sprognævnene i Norden ikke dækker alle former for sprogteknologi i Norden, udgør de i kraft af deres opgaveportefølje og sproglige opgaver en meget avanceret brugergruppe, og det er tydeligt at ønsket om at kunne bruge sprogteknologi generelt i de nordiske samfund og i sprognævnene i særdeleshed er steget betydeligt siden 2007. Endvidere

er nye kommunikationsformer kommet til, bl.a. var udviklingen i de sociale medier som Facebook og Twitter og i app-markedet til telefon og tabletcomputere kun lige begyndt da sprognævnenes kortlægning blev iværksat.

I takt med vedtagelser af sproglove i Sverige, Island og Grønland, og nye sprogpolitiske tiltag i de andre lande er der i varierende grad blevet investeret offentlige midler i udvikling af sprogteknologi og sproglige resurser. For tiden ser vi de største offentlige investeringer i termbanker og sprogbanker for norsk, svensk og finsk. Endvidere er det lykkedes at udvikle avanceret sprogteknologi for grønlandsk og samisk omfattende bl.a. stavekontroller, talesyntese og maskinoversættelse, og også i Island er der på det seneste opbygget et sprogteknologisk basisberedskab i form af sprogteknologiske korpusser, ordbøger og værktøjer til håndtering af islandsk. Danmark har udviklet værtøjer og indsamlet i sproglige resurser i DK-Clarin fra 2008-2011 (Halskov & Asmussen 2009), og der bliver løbende udviklet sprogteknologi i forbindelse med nye forskningsprojekter, fx inden for taleteknologi og semantik, fx et dansk Wordnet (Nimb 2006 og www.wordnet.dk), men der er ikke foretaget større nationale satsninger siden, og der foregår kun sporadisk koordinering af de eksisterende projekter og resurser.

Norden som sprogteknologisk underudviklet region?

Samtidig med at udviklingen for de større sprog accelererer stadig kraftigere, ser vi at mange basale problemer for de nordiske sprog stadig ikke er løst. Hvor man i de engelsktalende lande allerede er langt fremme med stemmestyrede applikationer, avanceret maskinoversættelse og semantisk fortolkning af tekst, er helt basale problemer, stadig ikke løst.

Der findes fortsat programmer hvor ŋ, ð, æ, ø, ö, å, fremstår som andre tegn eller omskrives til fx ae, oe, aa, og der er løbende problemer med at genfinde information fordi systemernes interne søgeprogrammer ikke er ordentligt gearet til de nordiske sprog. Selv om man har fundet en teknisk løsning på brugen af de særlige nordiske bogstaver i URL'er, er denne løsning ikke særlig udbredt, og da den ikke er blevet fulgt op af en tilsvarende løsning for e-mailadresser, ser man løbende især virksomheder og offentlige institutioner ændre stavningen af deres officielle navne til noget der er kompatible med det engelske alfabet. I Danmark skiftede byen Århus i 2012 således officielt til stavemåden Aarhus.

Selv simple applikationer som stavekontroller bliver ikke tilpasset de

nordiske sprog. Microsofts stavekontrol i Office 2010 foreslår fx som default at ord der ikke kan genkendes, deles i to, fx *garderobenicherne* foreslås skrevet *garderobe nicherne*, og for et simpelt ord som *dørhul* foreslås *dør hul*. Der findes endnu ingen undersøgelser af hvilke konsekvenser dette har for stavefærdigheden hos nordiske borgere, især børn og unge, men de er formentlig i det lange løb bedre tjent med ikke at få et forslag end at få et der er så misvisende.

Ser vi på kvaliteten af mere avanceret sprogteknologi, som fx maskinoversættelse, lader kvaliteten i fx Google Translate af de sprogpar som involverer nordiske sprog, fortsat meget tilbage at ønske, hvilken kan skyldes at der ikke er tilstrækkeligt mange parallelle tekster tilgængelige som involverer de nordiske sprog, sammenlignet med hvad der er tilgængeligt fx for de officielle UN-sprog som har dannet udgangspunktet for udviklingen af programmet (Och 2003).

Eksemplerne viser at der fra kommerciel side ikke satses på nordiske sprog, formentlig fordi markedet er for lille. Mulighederne for at sprogteknologien kan bidrage substantielt til at gøre Norden til en sproglig foregangsregion, er derfor kun til stede hvis forskning og udvikling inden for sprogteknologi får højere prioritet, og det tværnordiske samarbejde på området intensiveres.

Bedre sprogteknologi gennem samarbejde

Når det for de store softwareproducenter tydeligvis ikke er en god forretning at prioritere de nordiske sprog, er det helt afgørende at regeringerne i de enkelte lande er opmærksomme på disse problemer, og at de træder til med midler til at holde den sprogteknologiske udvikling i Norden på niveau med andre lande.

Samarbejdet kan med fordel foregå på flere niveauer, dels mellem sprognævn, forskere og virksomheder i de enkelte lande, dels mellem disse forskellige aktører på tværs af de nordiske lande. Endvidere bør repræsentanter for både nationale og nordiske offentlige institutioner deltage, da det i stigende grad også er den offentlige sektor som har brug for sprogteknologi. Da udviklingen på dette område går meget hurtigt, kan alle drage nytte af videndeling på tværs af sektorer og lande.

Et eksempel på dette er de danske kommuner som satser intensivt på at indføre taleteknologi som led i effektiviseringen af den offentlige sektor. Odense Kommune alene investerede i 2012 mere end 50 millioner kr. på udvikling af statistisk baseret taleteknologi. Mere end 2000 medar-

bejdere skulle begynde at tale til deres computer i hverdagen i stedet for at skrive breve, rapporter og indberetninger. Projektet var tæt på at kuldsejle da det viste sig at kvaliteten var alt for ringe, bl.a. fordi leverandørerne havde alt for lidt kendskab til dansk, og systemerne ikke var trænet på kommunale tekster. Kommunen skulle derfor hurtigt fremskaffe store tekstmængder for at virksomheden kunne gentræne systemerne og forbedre kvaliteten.

Fremskaffelsen af det sproglige råmateriale til træning af systemerne er dyrt og tidkrævende, men i modsætning til andre råstoffer kan materialet bruges igen og igen uden at blive slidt eller opbrugt. Der er derfor god økonomi i at sørge for at så meget af materialet kan genbruges, og det forudsætter afklaring af ophavsrettigheder og ejerforhold. Stort se alle nordiske kommuner vil i løbet af de kommende år foretage lignende investeringer som Odense Kommune. Kommunernes Landsforening har gennem sit eget selskab KOMBIT i samarbejde med DANCAST-centret ved CBS i København iværksat initiativer til at finde metoder til at sikre udbud af sprogteknologiske opgaver til en rimelig pris hvor de særlige krav til det sproglige råmateriale tilgodeses, og hvor kommunerne sikres mulighed for og rettigheder til at genbruge og dele det sproglige råmateriale (se også Juel Henrichsen i dette bind).

Brugen af nye input-metoder som talegenkendelse åbner for mange nye muligheder især i den offentlige sektor. I takt med at vi i stigende grad går over til borgerbetjening på nettet, vil borgere med læse-/skrivehandicap få problemer. Muligheden for talt input kunne være en løsning for mange. En del kommunale medarbejdere i socialsektoren har ligeledes svært ved at udtrykke sig korrekt på skrift i rapporter og indberetninger, og også her kan en talegrænseflade hjælpe. Opgaven med at sikre et korrekt og klart sprog lægges dermed i højere grad over på softwareprogrammerne – et område som de nordiske sprognævn derfor bør overvåge, således at de kan sikre at programmerne overholder de officielle standarder for skriftsproget.

Den digitale udvikling og den stigende mængde af skriftligt materiale på nettet og i diverse applikationer stiller imidlertid også store krav til sprognævnenes egen brug af sprogteknologi. Der er brug for software der kan hjælpe sprognævnene med deres løbende opgaver, fx

- Tekstanalyseværktøjer til analyse af aktuel sprogbrug (fx konkordansprogrammer, ordtrawlere m.m.)

- Orddatabaser til avanceret strukturering af ordbøger

- Programmer til udvikling af sproglige øvelser og tests

- Oversættelsesværktøjer mellem de nordiske sprog, især minoritetssprog, fx svensk-finsk, dansk-grønlandsk, norsk-samisk mv., og mellem de nordiske sprog og engelsk

- Værktøjer til kvalitetsanalyse af sprogteknologi, fx af stavekontrolprogrammer og af output fra talegenkendelse

Denne type software kan udvikles i samarbejde med virksomheder og forskningsinstitutioner, og også her vil der være meget at spare gennem videndeling og fælles nordisk programudvikling.

Sprognævnene kan som sagt også selv bidrage til at forbedre sprogteknologi ved at dele deres resurser med forskere og udviklere, fx

- Tekstsamlinger

- Talesprogsdata

- Ordbøger

- Grammatikker

- Andre typer af sproglig information, fx samlinger af nye ord og udtryk

Det ville være en stor fordel hvis dette skete systematisk og regelmæssigt, da man derved i højere grad kan sikre at de standarder som nævnene anbefaler, kommer ind i den software der udvikles.

Et hidtil overset forskningsområde er hvordan sprogteknologisk software kan holdes løbende ajour med den sproglige udvikling. Når et program trænes på en bestemt tekstsamling, vil det kun være i stand til at genkende og producere inden for de konstruktioner og det ordforråd som det er blevet trænet på, og hvis det ikke bliver opdateret, vil de fleste systemer blive forældet i løbet af 5-10 år. Sprognævnene har en særlig erfaring med at dokumentere forandringer i skriftsproget og bør også her

sørge for at dele deres viden og deres sprogresurser, fx nyordslister, mest muligt.

Til sidst kan sprognævnene bidrage til at øge kendskabet og udbredelsen af sprogteknologi og ikke mindst kendskabet til teknologiens begrænsninger og korrekte anvendelse.

Nævnene bør derfor fortsat

- opfordre beslutningstagere til at støtte udviklingen af sprogteknologi
- opfordre beslutningstagere til at gøre sprogresurser tilgængelige for sprogteknologi
- overbevise beslutningstagere om at det er vigtigt at der findes ekspertise inden for sprogteknologi for deres sprog

Det er helt afgørende at der på nationalt og på nordisk plan kontinuerligt iværksættes forskningsprogrammer og infrastrukturinitiativer der støtter udviklingen af sprogteknologi hvis de nordiske lande ikke skal sakke yderligere agterud på dette område.

Hvad angår sproglige rådata, bør man overveje at tænke en sprogteknologisk infrastruktur sammen med det nyligt reviderede PSI-direktiv fra EU (DIRECTIVE 2013/37/EU)[1] som fastlægger regler for genbrug af data i den offentlige sektor. Direktivet fastlægger at alle offentlige dokumenter skal kunne genbruges inden for direktivets rammer, at de skal gøres tilgængelige digitalt og forsynes med relevante metadata der følger åbne standarder, og at omkostningerne til dette skal holdes på minimumsniveau. Ved den seneste revidering i 2013 blev biblioteker, museer og arkiver ligeledes omfattet af direktivet, og det åbner yderligere mulighed for at inddrage relevante sproglige rådata.

Udvikling af sprogteknologisk software er ikke en opgave der kan overlades til almindelige it-ingeniører eller udenlandske softwarehuse. Det kræver både datalogisk indsigt og sproglig ekspertise – ikke blot en generel kompetence inden for kognition og lingvistik, men også en om-

1 http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information

fattende indsigt i de sprog som softwaren skal håndtere, helst som modersmålstalende.

Nordiske sprogteknologer og datalingvister udspringer af højtspecialiserede niceuddannelser som er meget sårbare over for de stigende effektiviseringskrav der stilles til universiteterne da der typisk kun uddannes en håndfuld i hvert land om året. I takt med at sprogteknologisk software bliver mere og mere udbredt, vil behovet for veluddannede sprogteknologer begynde at stige. Det er vigtigt i alle nordiske lande at man sørger for at der er stærke faglige miljøer på universiteterne der kan tiltrække nye sprogteknologer og uddanne dem på alle niveauer til alle nordiske sprog.

Konklusion

Sprogteknologi er vigtig for at sikre de nordiske sprogs status og brug i alle sammenhænge. Det påpeges allerede i den nordiske sprogdeklaration i 2006, men i lyset af den rivende udvikling på området burde anbefalingerne i sprogdeklarationen vedr. sprogteknologi revideres. De nordiske sprognævn har en forpligtelse til at sikre at sprogteknologi af høj kvalitet bliver udviklet for deres respektive sprog. Samarbejdet mellem forskere og udviklere inden for sprogteknologi og de nordiske sprognævn bør yderligere styrkes for at sikre løbende vidensudveksling. Beslutningstagere bør især have fokus på at etablere en permanent sprogteknologisk infrastruktur der sikrer indsamling og genbrug af sproglige rådata i form af sprogbanker, fx i tilknytning til implementeringen af EU's PSI-direktiv.

Sabine Kirchmeier-Andersen er sprogteknolog og direktør i Dansk Sprognævn.

Summary

Language technology is important for the status and the use of Nordic languages in all domains. Language institutions have an obligation to ensure that language technology is developed for their language. The Nordic language councils have worked eagerly to promote the knowledge of language technology and to bring together relevant players in the field in order to strengthen Nordic cooperation. The language council working

group for language technology in the Nordic countries (ASTIN) has held a number of successful conferences, seminars and workshops on various aspects of language technology. The article describes the results of ASTIN's work, and how the cooperation on language technology can be strengthened in order to meet the challenges of the information society of the future. The recommendations include the sharing of language data in a strong language technology infrastructure and a constant focus on the education of language technology experts with competence in Nordic languages.

Litteratur

Kirchmeier-Andersen, Sabine, 2011: Language Technology for Language Institutions. I: Stickel, Gerhard & Tamas Varadi (eds.): *Language, Languages and New Technologies. ICT in the Service of Languages*. Duisburg Papers on Research in Language and Culture. Vol. 87. Peter Lang, s. 21-32.

Crystal, David, 2000: *Language Death*. Cambridge. Cambridge University Press.

Koskenniemi, Kimmo, Lindén, K., Nordgård, T., 2007: *Språkvis – Språkteknologisk vismansrapport. En utvidgad sammanfattning av The Nordic Countries – A leading region in Language Technology*. Helsingfors.

EU (DIRECTIVE 2013/37/EU)

Deklaration om nordisk sprogpolitik 2006. Nordisk Ministerråd.

Krauwer, Steven, 1998: *ELSNET and ELRA: A common past and a common future*. In: ELRA Newsletter Vol. 3 N. 2.

Nimb, Sanni & Hartvig Sørensen, Nicolai, 2006: DANNET – et leksikalsk semantisk wordnet for dansk. I: Peter Widell og Ulf Dalvad Berthelsen (udg.): *11. Møde om Udforskningen af Dansk Sprog.* Århus.

Juel Henrichsen, Peter, 2014: Tale er Guld. Om talegenkendelse i de danske kommuners tjeneste. I: Språk i Norden 2014.

Langgård, Per, 2014: Selvfølgelig snakker vaskemaskiner grønlandsk i fremtiden. I: Språk i Norden 2014.

Moshagen, Sjur, 2014: Status for samisk språkteknologi. I: Språk i Norden 2014.

Och, Franz Josef, 2005: Statistical Machine Translation: Foundations and Recent Advances. In: *Proceedings of the Tenth Machine Translation Summit*, Phuket, Thailand 2006.

Halskov, Jakob & Jørg Asmussen, 2009: Compiling, annotating and publishing corpora in DK-CLARIN, the Danish incarnation of the pan-European initiative for a common research infrastructure. In: *Work-in-progress report. Corpus Linguistics*. Liverpool University.