

# SPRÅKRÅDET

## NOTAT

**Fra/frå:** Torbjørg Breivik, Språkrådet

**Til:** Kultur- og kirkedepartementet

**Dato:** 12.08.2009

**Sak:** Norsk språkbank – oppdatering av opplysninger og vurderinger

### **Innledning**

Notatet er en gjennomgang og en utdyping av problemstillingene Kultur- og kirkedepartementet ber om tilbakemelding på i brev til Språkrådet 29. mai 2006.

Vi starter med å vise noen av mulighetene språkteknologien gir oss. Videre kommenterer vi planen for en norsk språkbank fra 2002, og vurderer hvilken verdi språkressursene i boet etter Nordisk Språkteknologi Holding AS på Voss kan ha i en framtidig språkbank. Situasjonen i Norge og de nordiske landene slik den er i dag, gjennomgås, og særlig kommenteres utredningen om en nordisk språkbank. Vi har også tatt med litt om hvor infiltrert hverdagen vår alt er av språkteknologiske tjenester og produkt, nettopp for å synliggjøre at en satsing på norsk språkteknologi er konkret, nyttig og verdifull.

### **Språkteknologi gir oss uante muligheter!**

Vi står ved et veivalg i norsk språkpolitikk. Språkteknologien kommer vi ikke utenom uansett hvilken retning vi velger framover. Det har vært diskusjoner om norsk vil være et utdødd språk om hundre år. Det har vært diskusjoner om bruk av engelsk eller norsk på ulike samfunnsområder. Språkteknologien aktualiserer begge problemstillingene. Mange språkteknologiske produkt finnes i dag kun på engelsk, særlig de mest avanserte. Norge betrakter seg som et kunnskapssamfunn som utvikler og bruker avansert teknologi der vi kan. Det er ikke uvanlig å høre uttalelser fra høyt politisk hold om at Norge skal være i verdenstoppen på område x og y ved å satse på innovasjon og utvikling. Norske språkteknologiske fagmiljø holder høyt internasjonalt nivå, og vi må bruke mulighetene det gir oss. Det står mer på spill enn en mulig bærekraftig kunnskapsindustri, det gjelder hele den norske språk- og kulturarven.

### **Språkteknologi er alt hverdagen vår**

Hvor mange vet at det er språkteknologien som gjør det mulig for oss å benytte mobilens intelligente ordbok til å skrive raskere meldinger? Og hvor mange vet at det er språkteknologien som gjør det mulig for et automatisk sentralbord å forstå hva vi

# SPRÅKRÅDET

ber om, eller som (ofte) spør om igjen for å forsikre seg om at spørsmålet er riktig forstått? Hvor mange vet at stave- og grammatikk kontrollene i tekstbehandlingsprogrammene vi har, er språkteknologiske produkt? Og hvem har ikke irritert seg over alt stavekontrollen "godkjenner" av rariteter som slett ikke er i overensstemmelse med norsk rettskriving? Språkteknologi er i bruk i bilenes navigasjonssystem, i teksting av film, oversetting av tekst, til styring av vaskemaskiner osv. Når vi søker på nettet, finnes det i dag intelligente søkesystem som fortolker det vi søker etter, for å gi oss et best mulig svar. Det finnes program som støtter oversetting av tekst, som kan lage sammendrag av tekst og rangere relevante data fra dokument. Språkteknologi bidrar til økt demokratisk deltakelse i samfunnet bl.a. gjennom å gjøre informasjon og tjenester tilgjengelig for alle. Språkteknologi er for alle uansett språk, kjønn, sosial og etnisk tilhørighet, psykisk og fysisk funksjonsnivå, språklig og teknisk kompetanse, utdanningsnivå og arbeidsområde. Språkteknologien bidrar til å styrke det norske språket og den norske kulturen i en global verden.

Språkteknologi gjør det mulig å kommunisere uavhengig av om brukeren kun kan snakke, høre eller skrive. Funksjonshemmede som ikke kan skrive selv, kan snakke til en pc som oversetter talen til tekst. Pc-en styres med stemmen i stedet for at en skriver kommandoene, og tekst kan sendes til en mottaker som kan velge om han/hun vil lese teksten eller få den opplest. Offentlig informasjon er tilgjengelig på Internett, og "24-timers-staten" realiserer at skjemaer og informasjon tilrettelegges for interaktiv bruk. Brukeren velger når han/hun vil benytte informasjonen, fyller ut og sender ulike søknader og skjemaer, og språkteknologien gir brukeren valget mellom å lese og lytte, og så f.eks. fyller ut skjemaer ved hjelp av tastaturet eller stemmen. Det er ulike deler av språkteknologien som tas i bruk for å kunne tilby disse mulighetene. Vi har behov for alle, og vi har behov for dem på norsk.

Alle skjønner at i et samfunn med analfabeter vil maskiner og hjelpemidler med skriftlige brukerveiledninger ikke fungere. *Språkteknologien tas i bruk for at ikke analfabetisme skal hindre et samfunn fra å følge med i den teknologiske utviklinga.* I et høyteknologisk samfunn er det mange arbeidsoppgaver som kan forenkles og effektiviseres ved hjelp av språkteknologi, det er bare fantasien og tilgang til norskspråklige versjoner som setter stopper for norske borgeres muligheter til å nyttiggjøre seg dem. Teknologien setter fortsatt visse begrensninger, særlig når det gjelder bruk av naturlig tale.

## Tidligere utredninger og planer

*Informasjons- og kommunikasjonsteknologien er i rask utvikling.* Prosjektplanen "Samling og tilgjengeleggjering av norske språkteknologiressursar" ble skrevet i 2002, og mye har skjedd med teknologien på tre–fire år. Når man i "Norsk i hundre!" på side 122-124 kommer med uttalelser som dem KKD refererer i brevet til Språkrådet datert 29.05.2006, kan den raske teknologiutviklinga være grunnen.

Prosjektplanen fra 2002 peker på muligheten for å ta inn noe av driftsutgiftene for en norsk språkbank i form av brukeravgifter. Sitatet som KKD har hentet fra "Norsk i hundre" (NiH), må være tatt ut av sammenhengen. I prosjektplanen fra 2002, kapittel 6.3, side 35 står det: "Brukarane, dvs. industri og forskning, må betale for bruken av innhaldet i samlinga, og noko av kostnadene med drifta bør etter kvart kunne dekkjast med slike vederlag." Dette rører ikke ved det faktum at **etableringen** av

# SPRÅKRÅDET

språkbanken er den store utgiften. Rapporten fra 2002 hadde med kalkyler basert på bidrag fra forskning og næringsliv i form av innsamlede ressurser, og det skulle være en viss betaling for (kommersiell) bruk. I dag, som i 2002, er det vanskelig å vurdere eksakt hvor stor andel av driftskostnadene for språkbanken som vil kunne dekkes gjennom brukeravgifter. Betaling for bruk kan dekke driftsutgifter og noe utvikling/vedlikehold. Etableringskostnadene må hovedsakelig være offentlige. Forutsetningene har ikke endret seg fra 2002 på dette punktet, og "Norsk i hundre!" baserer seg på de samme forutsetningene. "Norsk i hundre!" er opptatt av å peke på de økonomiske gevinstene som kan høstes – både i form av å etablere en språkteknologisk industri i Norge, men aller mest hvilke besparelser man får ved å ta i bruk språkteknologi (jf. eksemplet om å spare 10 % av dikteringskostnadene ved norske sykehus, (NiH, s. 9). Språkbanken vil være avhengig av driftstilskudd over tid for å bli realisert og komme i ordinær drift. Hvordan et driftstilskudd gis, vil avhenge av hvilken organisasjonsmodell man velger for språkbanken. Etableres språkbanken som en stiftelse med mulighet for å ta imot gaver/donasjoner fra private aktører i tillegg til midler fra det offentlige, kan man fordele kostnadene på mange og lette finansieringen. I land som USA, Italia, Frankrike, Tyskland og Nederland har man lagt en slik modell til grunn, basert på avtaler mellom stat, forskningsråd og store private aktører i markedet. En språkbank er ikke et prosjekt som har noen sluttdato, men det er etableringen som er kostnads- og ressurskrevende.

Språkrådet ser ikke konturene av en utvikling som skulle tilsi til at en infrastruktur for språkteknologi (en språkbank) har noe stort inntjeningspotensial slik vi ser framtida for oss i dag. Hovedpoenget med å utvikle norsk språkteknologi er besparelsene man får på forskjellige områder ved å ta i bruk språkteknologi, best illustrert ved eksemplet fra sykehusverdenen (10 % reduksjon av dikteringskostnadene). Det koster omtrent det samme å etablere og vedlikeholde den nødvendige infrastrukturen for språkteknologi uansett hvilket språk vi tar for oss. Norsk har få brukere å fordele kostnadene på, et lite marked å selge i, og det kan derfor virke uforholdsmessig dyrt for oss.

Den engelskspråklige verden utgjør et stort og kjøpesterkt marked med gode muligheter for inntjening. I USA er det tradisjon for å lage konsortier der offentlige og private interesser går sammen om å finansiere et prosjekt av denne typen. Men selv med et så stort og kjøpekraftig marked som det amerikanske ble infrastrukturen for språkteknologien finansiert med offentlige midler: forsvars- og forskningsdepartementene (NASA og ARPA) sørget for midler til innsamling, bearbeiding og tilrettelegging av språkressursene som nå er allment tilgjengelig for forskning og industri gjennom basen Linguistic Data Consortium (LDC). LDC er et konsortium dannet av universiteter, private firmaer og statlige forskningsinstitusjoner, og administrert av University of Pennsylvania ([www ldc.upenn.edu](http://www ldc.upenn.edu)). LDC dekker mye av sine driftskostnader gjennom medlems- og brukeravgifter, men får fortsatt offentlige midler via National Science Foundation. Den europeiske språkbasen i Paris, ELDA, er et selvstendig juridisk firma (pga. fransk lovgiving). Den tar imot språkressurser andre har finansiert og laget, og formidler gjenbruk av disse gjennom en medlemsorganisasjon der man betaler for bruken (gradert betaling: minst for forskningsformål, mest for kommersiell bruk). ELDA finansierer ikke nyinnsamling eller bearbeiding av språkressurser, det er kun et distribusjonsorgan.

# SPRÅKRÅDET

Norge har et høyt teknologisk kunnskapsnivå. På mange områder er ny og tilgjengelig teknologi tatt i bruk, og i norsk politikk hevdes det stadig at det er viktig å satse mer på ny teknologi. Skolen er en sentral arena for å oppnå målene man har satt seg for kunnskapssamfunnet. Utdanningsdepartementet er opptatt av digital kompetanse. Språket er et grunnleggende element i læringsprosesser og utvikling av kompetanse, også den digitale. Det er vanskelig å diskutere hva kunnskap er uten å gi det et språklig uttrykk. Få stiller spørsmål ved at undervisningsspråket i grunnskolen er norsk. Når man i undervisningen i flere og flere fag tar i bruk digitale læremidler, sier det seg selv at læremidlene må foreligge på norsk. Men det gjør de i liten grad i dag! Som nevnt ovenfor blir språkteknologi mer og mer en del av andre produkt, og derfor er det viktig at den blir gjort tilgjengelig på norsk. For at språkteknologien virkelig skal bli et nyttig verktøy og hjelpemiddel, må den basere seg på kunnskap og kompetanse om det norske språket: bokmål, nynorsk og talemålsvariantene.

Språkrådet ser språkteknologien som en naturlig del av kunnskapssamfunnet, og Språkrådets overordnede mål er at norsk fortsatt skal være det foretrukne språket i Norge på dette området som på alle andre samfunnsområder. Å etablere en infrastruktur for norsk språkteknologi er strategisk viktig og nødvendig om Norge fortsatt skal ha norsk som samfunnsbærende språk. Språkteknologien blir en del av det å ivareta norsk kultur fordi den trenger inn på alle samfunnsområder.

Språkrådet vurderer etablering av en norsk språkbank på samme måte som i 2002. Teknologien er forbedret. En viss internasjonal standardisering har skjedd og kan redusere utgiftene til nyinnsamling og tilrettelegging en del. Kravene vi må stille til omfang og innhold, har ikke endret seg i årene som har gått. Det har skjedd mest når det gjelder å nyttiggjøre seg behandlet tale og lyd (neste nivå i utviklingskjeden), men kravet til infrastruktur er ikke vesentlig endret. Ressurssituasjonen har heller ikke endret seg i løpet av de fire årene: Språkrådet mener fortsatt at universitets- og forskningsmiljøene må løfte dette prosjektet sammen, også sammen med relevante aktører i industrien. Det er samlet inn en del nye ressurser ved universitetene, bl.a. for prosjektene i NFRs KUNSTI-program og ellers. Disse ressursene kan endre bildet når det gjelder behovet for nyinnsamling, og en **oppdatert kartlegging** av hvilke ressurser som finnes og hvilke områder som ikke er dekket, må ha **førsteprioritet** når man går i gang med prosjektet.

Det vil være en stor fordel om private kan innlemmes i prosjektet slik man forutsatte i planen fra 2002. I hvilken grad private aktører kan og vil bidra, avhenger av hvilken organisasjonsmodell man velger for språkbanken:

- statlig aksjeselskap eller
- offentlig stiftelse

***Enten man velger den ene eller andre organisasjonsmodellen, må styret settes sammen slik at det representerer de språkteknologiske fagmiljøene i Norge, inklusive industrien.***

Man må også sikre at språkbanken forblir i offentlig eie, og lage vedtekter som hindrer at private investorer eller andre interesser kan handle med språkbanken etter eget for godtbeholdende. Språkbanken må ikke utsettes for skiftende eierinteresser

# SPRÅKRÅDET

slik vi i dag ser en del mediebedrifter blir. Industriforetak kan bidra med språkressurser og/eller med verktøy til bearbeiding og formidling av innholdet i språkbanken. Som et minimum bør det kunne åpnes for donasjoner og gaver fra private aktører. All innsats koster, og infrastrukturen, selve etableringen av språkbanken, er kostnadskrevende! Velger man en organisasjonsmodell med åpning for deltakelse/donasjoner fra private aktører, vil man ikke nødvendigvis måtte ha 100 mill. kr over fem år i *friske* midler for å realisere prosjektet. Deler av språkbanken kan finansieres ved at språkressurser doneres/gis til språkbanken. Man kan også se for seg at man finansierer innhold i språkbanken ved å spesifisere en verdi på ressurser språkbanken mottar, og at verdien blir stående som et innskudd i språkbanken uten at penger overføres mellom leverandør og mottaker. Transaksjonene må hjemles i juridiske avtaler som gjør ressursene tilgjengelige for andre. Det er også mulig å tenke seg at en leverandør kan ta ut andre språkressurser for tilsvarende verdi som det man har levert inn. Ordninger som dette er for øvrig skissert i prosjektplanen fra 2002.

Når det gjelder utgiftene, kan den totale kostnadsrammen justeres litt ned. Forutsetningen er at ressursene fra NST-boet frikjøpes og legges inn i prosjektet (se nedenfor). En konkret gjennomgang av eksisterende ressurser hos universitetene og hos interesserte industriaktører (f.eks. Telenor) vil avklare hvor mye og hvilke språkressurser som ellers kan være aktuelt for språkbanken. Denne kartleggingen vil avdekke hvilken type ressurser vi ikke har, og som så må samles inn fra grunnen av. En stor del av de ressursene vi kartla i 2002, var innsamlet til forskningsformål. Opphavsrettslige avtaler stenger for at de kan innlemmes i en språkbank. Flere fagmiljø har etter 2002 samlet inn språkressurser til eget bruk. Noen av dem har vært oppmerksom på problemstillingene knyttet til opphavsrett, og kan ha sørget for avtaler som gjør at ressursene kan innlemmes i en framtidig språkbank. Men her må det gjennomføres en ny kartlegging av type ressurser, rettigheter, omfang og kvalitet.

Framdriftsplanen fra 2002 la til grunn omfanget av kvalifisert, tilgjengelig arbeidskraft mer enn hvor mye tid det ideelt sett tar å etablere et tilfredsstillende minimum av innhold i en språkbank. De aktuelle fagmiljøene i Norge er fortsatt små, og kapasiteten i utdanningen har ikke økt på de aktuelle fagområdene. Fagmiljøet som var på Voss, er splittet. Det betyr at framdriften vil avhenge av kapasitet til å ta denne typen oppgaver (oppdrag) i de eksisterende miljøene. Det må utarbeides en ny og prioritert framdriftsplan i nært samarbeid med fagmiljøene.

Oppgaver (forutsatt at det foreligger vedtak om opprettelse av en norsk språkbank og midler til finansiering):

1. Opprette et interimsstyre som kan forestå oppstart, planlegging og prioritering. Styret må ha representanter fra forskning og industri, eventuelt også fra departementalt/politisk hold.
2. Det må tilsettes en prosjektleder som bl.a.
  - a. utarbeider en revidert prosjektplan med nye kostnadskalkyler
  - b. leder prosjektet/delprosjektene og følger opp styrets prioriteringer
  - c. kartlegger aktuelle ressurser: type, omfang, grad av tilrettelegging for gjenbruk osv.
  - d. avklarer rettigheter knyttet til dem
  - e. sørger for en nøytral, faglig gjennomgang av aktuelle ressurser
  - f. sørger for maler for avtaler angående ressursene (kjøp og salg)

# SPRÅKRÅDET

- g. inngår avtaler om frikjøp, mottak, plassering og tilgang til ressursene i språkbanken
- h. sørger for infrastruktur for selve språkbanken (innkjøp av maskiner, programvare, opprette nettside, samarbeid nasjonalt og internasjonalt osv.)
- i. tar hånd om kjøp, salg og drift av språkbanken

## Eksisterende språkressurser

*Restene etter Nordisk Språkteknologi Holding AS (NST) på Voss:*

I desember 2005 la Språkrådet fram to rapporter om språkressursene som finnes i databasene i boet etter NST. Den ene var en gjennomgang av innholdet i basene, og den andre en gjennomgang av de juridiske avtalene knyttet til de aktuelle delene av innholdet. De akustiske basene viste seg å være i god stand, holder høy kvalitet og er velegnet for gjenbruk. Basene omfatter opptak av norsk, svensk og dansk tale. Innholdet er bearbeidet systematisk, følger samme mal og internasjonale standarder er benyttet i den grad det fantes. De juridiske avtalene gir NST (og nå boet) rettigheter som tilsier at ressursene kan selges videre. Ved å løse ut disse ressursene fra konkursboet vil Norge ha en svært god start på en norsk språkbank. Konkursboet har krevd 4 mill. NOK totalt for språkressursene inkludert noen nesten ferdige produkt. Ressursene representerer mye arbeid for mange personer med ulik faglig kompetanse. Det er vanskelig å anslå nøyaktig hvor mye det vil koste å samle inn, merke og gjøre tilgjengelig tilsvarende mengde i dag, men et forsiktig anslag tilsier at det kan raskt dreie seg om fem–seks ganger så store kostnader.

Etter Språkrådets vurdering vil vi spare store beløp ved å kjøpe språkressursene fra konkursboet etter NST. Tale er noe av det dyreste man samler inn, bearbeider og tilrettelegger for språkteknologisk (gjen)bruk, og her har vi mulighet for å få en god del for en akseptabel pris.

Ressursene omfatter dansk og svensk i tillegg til norsk, noe som gjør at de kan være aktuelle i forbindelse med en nordisk språkbank.

En språkbank må ha andre taleressurser enn dem som er nevnt over, men ved å innlemme de eksisterende ressursene i banken kan man spare en hel del arbeid fordi dette er bearbeidet materiale tilrettelagt for bruk. Bearbeiding og tilrettelegging for gjenbruk av språkressurser forutsetter tilgang på kvalifisert arbeidskraft, krever mye tid og er kostbart. Norge, som de andre nordiske landene, har begrenset antall fagpersoner med den nødvendige tverrfaglige kompetansen i lingvistikk og teknologi som kreves for å utføre slike oppgaver. Hver lyd skal merkes, hvert ord skal merkes, grammatiske opplysninger skal være med, selve opptaket skal dokumenteres med sjanger, alder, kjønn, opptakssituasjon osv.

## Språkressurser – etablering og samarbeid

Her er situasjonen omtrent som i 2002 når det gjelder hvilke miljø som kan være aktuelle bidragsytere til språkbanken. Det som har skjedd i tiden fra prosjektplanen ble skrevet og til nå, er at miljøene har utvidet sine ressurssamlinger. De har fått

# SPRÅKRÅDET

flere, større og andre korpus i tillegg til dem som er listet opp i oversikten fra 2002, og som før er ressursene sannsynligvis bare tilgjengelig for forskningsformål. De språkteknologiske forsknings- og utviklingsmiljøene i Norge er ikke store.

Da Språkrådet fikk oppdraget med å utarbeide en prosjektplan for språkbanken i 2002, erfarte vi at fagmiljøene var lette å få med i arbeidet, at de ikke bare så det som formålstjenlig å samarbeide, men at de selv uttrykte et sterkt ønske om samarbeid. Prosjektgruppa hadde representanter for universitetene, fra Forskningsrådet og fra flere private aktører (Telenor, NST, CognIT). I tillegg ble en referansegruppe med representasjon fra flere aktører og en del mulige leverandører av språkressurser konsultert undervegs. Språkrådet tar ikke for sterkt i når vi hevder at fagmiljøene sto, og fortsatt står, samlet bak et uttalt behov for og ønske om at språkbanken må etableres snarest. Språkrådet har stadig henvendelser om hvor langt prosjektet har kommet og når man kan forvente å få levert ressurser fra språkbanken.

De språkteknologiske fagmiljøene er konsentrert til tre steder i landet: Oslo, Bergen og Trondheim (og en viss aktivitet knyttet til samisk i Tromsø). Det er et visst samarbeid mellom de ulike stedene, i alle fall innenfor universitetene. Noen store private forsknings- og utviklingsmiljø (f.eks. SINTEF og Telenor) samarbeider med universitetsmiljøene. Universitetsmiljøene utfyller hverandre når det gjelder områdene innenfor språkteknologien. I Oslo arbeider man mest med maskinoversetting, digitale ordbøker og noe tekstlingvistikk. I Bergen er det særlig store tekstsamlinger, en del arbeid med datalingvistikk og en del terminologi som står sterkt. Miljøet i Trondheim er mest kjent for arbeidet med tale og lyd, men også her arbeider man med datalingvistikk.

Den språkteknologiske kompetansen er høy selv om miljøene er små. Det gir Norge et godt grunnlag for å utvikle seg til å bli blant de ledende i verden om det satses tilstrekkelig på dette området. Norge har et høyt utdanningsnivå i befolkningen, stor pc-tetthet, og ny teknologi er med stort hell tatt i bruk på mange samfunnsområder. Det gir et godt utgangspunkt for et avansert kunnskapssamfunn, og i dette landskapet har norsk språkteknologi en selvfølgelig plass. Det er derfor viktig at man legger forholdene til rette for å utnytte de mulighetene vi har. Det koster like mye å utvikle språkteknologi for/på norsk som for/på engelsk, og samarbeid mellom de ulike aktørene, offentlige som private, er nøkkelen.

Forskningsrådet brukte om lag 70 mill. kr til det store språkteknologiske programmet KUNSTI. Prosjektene som fikk tildelt midler, var store og forutsatte at flere fagmiljø samarbeidet, og et av de svært positive resultatene er økt samarbeid og bedre oversikt over hverandres sterke sider i tilgrensende fagområder som man selv arbeider med. Flere av prosjektene i KUNSTI måtte bruke tid, krefter og penger på innsamling og bearbeiding av nødvendig infrastruktur (språkressurser) før de kunne gå i gang med selve prosjektene. Dette ga mindre ressurser til forskningsarbeidet, som igjen førte til justering av forskningsfokus og ambisjonsnivå for flere prosjekt. Forskningsrådet måtte justere sine ambisjoner for programmet bl.a. fordi de hadde trodd at språkbanken ville komme i gang samtidig som de startet KUNSTI. Disse prosjektene ble ikke koordinert, og det er viktig for alle at en slik situasjon ikke oppstår på nytt.

# SPRÅKRÅDET

## Nordisk

De nordiske landene har valgt ulike strategier når det gjelder satsing på språkteknologi.

*Island* utarbeidet og vedtok sin strategi og handlingsplan for islandsk språkteknologi for fem år siden. En inspirasjonskilde for dem var Språkrådets handlingsplan for norsk språk og IKT. Det er utdanningstilbud på universitetsnivå, og det er opprettet to større statlige fond for å finansiere prosjekt og språkbaser. Det finnes språkbaser der innholdet er tilrettelagt for språkteknologisk bruk. Basene er tilgjengelig både for forskningsformål og for private firmaer som utvikler språkteknologiske produkter. Resultatene har kommet i form av mer forskning på området og økt interesse for fagområdet, bl.a. ved at private firmaer leverer språkteknologiske produkter og tjenester på islandsk. Island har ikke etablert noen statlig språkbank.

I *Danmark* har Forskningsrådet for Kultur og Kommunikasjon utarbeidet en strategi for dansk språkteknologi (2004) på bestilling fra Statens Humanistiske Forskningsråd. I utredningen finnes et kostnadsoverslag for en dansk språkbank med forslag til innhold og hvor mye hver del koster. Totalsummen er litt lavere enn de norske beregningene fra 2002. Utredningen viser til anbefalinger fra Forskningsministeriets arbeidsgruppe for IT på dansk fra 2001 som foreslo å opprette en dansk språkbank. Ifølge utredningen fra 2004 arbeider Ministeriet for Videnskap, Teknologi og Udvikling med å konkretisere den danske språkbanken, hvilke oppgaver den skal ha, hvilke formål den skal støtte osv. ("Strategisk satsing på dansk sprogteknologi", Forskningsrådet for Kultur og Kommunikasjon, 2004, s. 82–83).

Det finnes flere språkteknologisk tilrettelagte ressurser i Danmark, og noen av dem er tilgjengelig for andre enn forskere, men det meste er der, som i Norge, bare for forskere og forskningsmiljøene. Det danske Kulturministeriet har imidlertid fulgt opp satsingen på området med å finansiere konkrete prosjekt som utvikling av en språkteknologisk ordbok, utvikling av automatisk teksting av tale på TV, midler til utvikling av hjelpemidler for handikappede og ikke minst penger til å utrede det de kaller "Dansk Sprogteknisk Servicecenter".

Center for Sprogteknologi har utviklet en språkteknologisk ordbok til bruk i språkteknologiske produkter og tjenester. Senteret arbeider mye med underlag for maskinstøttet oversettelse for språk i EU, og de deltar i, og leder, flere EU-prosjekt i språkteknologi. De har i mange år tatt til orde for å få en dansk språkbank for å samle eksisterende språkressurser og gjøre dem tilgjengelig for dem som trenger dem. I Danmark, som i Norge og Sverige, er fagmiljøene i språkteknologi spredt: Københavns Universitet, Handelshøjskolen i København, Aalborg Universitet og Syddansk Universitet.

*Sverige* har flere språkbanker, men bare for forskningsformål. Innholdet i dem varierer, men ingen av dem dekker alle områdene for en nasjonal språkbank. Vetenskapsrådet arbeider nå med en såkalt "roadmap" der behovet for infrastruktur for svensk forskning de kommende ti årene kartlegges. Språkteknologi nevnes der som et viktig område med stort behov for utvikling av infrastrukturen. Sentrale aktører i svensk språkteknologi (språkbanken i Göteborg, forskerskolen GSLT og

# SPRÅKRÅDET

Språkrådet) har gått sammen om en søknad til Vetenskapsrådet om midler til å etablere en språkbank. De har også søkt om midler til å utarbeide en plan for hvordan man kan utvikle og gjøre tilgjengelig nødvendige ressurser til språkbanken. Flere universitet og forskningsinstitusjoner enn dem som står bak søknaden til Vetenskapsrådet, er med i diskusjonene om å etablere en språkbank. De ønsker at innholdet i den svenske språkbanken skal kunne brukes både av forskningsmiljøene og av private firmaer som vil utvikle språkteknologiske produkt. Tankegangen er den samme som i Norge: Vi har begrensede ressurser, små markeder og dette er dyrt, derfor må man i felleskap lage en ressurs-samling som alle kan bruke.

Også i Sverige er språkteknologimiljøene spredt på flere steder med universitetene i Uppsala, Stockholm, Lund, Linköping, Göteborg, høgskolen i Växjö, Kungliga tekniska högskolan i Stockholm og Chalmers tekniske universitet som de mest sentrale.

Sitatet fra "Bästa språket.." i brevet fra KKD presiserer at det nye statlige svenske språkorganet (Språkrådet) ikke selv kan gjennomføre arbeidet med å etablere en språkbank. Men Språkrådet bør ha kompetanse til å inventere, ha oversikt og initiere nødvendige samarbeidsprosjekt som kan bidra til å realisere en språkbank. Dette tilsvarer situasjonen i Språkrådet i Norge der man alt har denne typen kompetanse.

*Finland* har en språkbank for forskningsbruk. Firmaer som har bruk for språkressurser i sin produktutvikling, er avhengig av å samle inn disse selv. Det foreligger ikke planer om etablering av noen finsk eller finsk-svensk språkbank finansiert med statlige midler. Fagmiljøene er i stor grad konsentrert i og rundt Helsingfors med universitetet som en sentral drivkraft. Finland har noen språkteknologiske bedrifter som er underleverandører til store konsern som Microsoft og Nokia. Det er et paradoks at det er et finsk firma som leverer den norske stave- og grammatikkontrollen til MS Word, inklusive nynorsk stavekontroll!

*Færøyene* har en samling akustiske ressurser som er brukt til utvikling av en talesyntese for færøysk. Denne befinner seg fysisk hos universitetet i Stockholm fordi vedkommende som utviklet den færøyske talesyntesen, holder til der. Ressursene er ikke åpent tilgjengelige, men det er ikke noe problem å få tilgang til å bruke dem. Prosjektet ble finansiert av det færøyske hjemmestyret og det færøyske blindedeforbundet.

*Grønland* har ikke selv den nødvendige kompetansen verken på universitetet eller i det private næringslivet til å sette i gang større språkteknologiske prosjekt. Den grønlandske språknemnda arbeider med en språkteknologisk ordbok for inuittisk som etter planen skal være ferdig rundt årsskiftet 2006–2007. Arbeidet er finansiert av det grønlandske hjemmestyret.

## En nordisk språkbank

Nordens språkråd har tatt initiativ til å utarbeide en vismannsrapport med strategi for en nordisk språkteknologipolitikk. Rapporten skal leveres Nordisk råd, og skal stake ut en felles nordisk strategi for språkteknologiområdet og gi anbefalinger for en nordisk satsing på språkteknologien de neste ti årene. Rapporten forutsetter at hvert

# SPRÅKRÅDET

av landene får på plass den nødvendige underliggende infrastrukturen for en slik satsing, bl.a. ved å etablere nasjonale språkbanker. En nordisk språkbank kan fungere på to måter:

- de nasjonale språkbankene tilbyr sine ressurser til andre gjennom en felles nettportal
- man kan etablere en felles administrasjon/distribusjonsenhet for den nordiske språkbanken som også tar hånd om de respektive nasjonale språkressursene.

Den siste løsningen krever et mer formelt juridisk samarbeid på nordisk nivå. Den første løsningen fordrer bare samarbeid om en nettside, mens alt praktisk arbeid inklusive oppdateringer, drift og vedlikehold utføres i de respektive språkbankene.

**Et samarbeid mellom landene i Norden vil, uavhengig av hvor tett det er, kunne bidra til å redusere deler av utgiftene til utvikling av verktøy fordi flere språkbanker kan nyttiggjøre seg dem.**

Vismannsrapporten skal behandles i Nordens språkråd i september 2006, og vil deretter bli overlevert Nordisk råd for behandling.

Vismannsrapporten viser til det europeiske konseptet Basic Language Resource Kit (forkortet BLARK), som spesifiserer hva en språkbank må inneholde av typer ressurser, omfang, kvalitet, standarder som bør følges, og hvilke verktøy som må finnes for at en språkbank skal være nyttig for språkteknologisk forskning og utvikling. Den norske språkbankrapporten fra 2002 la spesifikasjonene i BLARK til grunn for sin vurdering av innhold og omfang. BLARK var ganske nytt i 2002, men har etter hvert blitt den standarden man har lagt til grunn ved etablering av språkbanker i Europa. (I *Proceedings LREC 2002* (Third International Conference on Language Resources and Evaluation), Las Palmas de Gran Canaria, Spain, nettside: [www.lrec-conf.org/lrec2002](http://www.lrec-conf.org/lrec2002))

Vismannsrapporten har ingen spesifisert kostnadsoversikt over hva en nordisk språkbank vil koste, men anslår kostnadene per språk. Den peker på at det vil lønne seg for alle landene at de gjør innsamlingen omtrent samtidig, og følger samme standardisering når det gjelder lagring, merking, kontrakter osv. Rapporten anslår kostnadene per språk til 10–15 mill. euro for en tekstsamling, 10–20 mill. euro for en talespråksamling og 10 mill. euro for utvikling av et leksikon. I tillegg kommer utstyr, programvare, administrasjon og utarbeiding av kontrakter for kjøp og salg av ressurser. Det er uklart hva kostnadsoverslagene baserer kalkylene på, men kostnadskalkylene for den norske språkbanken går ut fra etablering av en minimumsløsning med helt grunnleggende ressurser og nødvendige verktøy for behandling av dem. Vismannsrapporten (utkast per 07.08.2006) kan leses på <http://forums.csc.fi/kitwiki/pilot/view/Main/LTExpertPanelReport?cover=print.pattern>.

Språknemndenes arbeidsgruppe for språkteknologi sto for planlegging og gjennomføring av en nordisk arbeidskonferanse i Finland i april 2005 om stavekontrollprogram (en ny er planlagt i oktober 2006). Konferansen var svært praktisk rettet og samlet deltakere fra forskning, industri og språknemnder. Flere av deltakerne fra industrien uttalte at de ikke bare kunne tenke seg, men vil se det som

# SPRÅKRÅDET

en stor fordel å betrakte Norden som ett marked. Noen av firmaene utvikler flere språkteknologiske produkter, og er interessert i å selge i hele det nordiske markedet. De uttalte at de vil ha stor nytte av nasjonale språkbanker, og at det ville være enklere om de kunne få alt fra ett sted: fra en nordisk språkbank.

Nordens språkråd har under det danske og det norske lederskapet de to siste årene prioritert språkteknologi og tatt initiativ til flere språkteknologiske prosjekt enn Vismannsrapporten.

En flerspråklig søkemotor og en nordisk nettdordbok er to av prosjektene som går mot sin avslutning. Målet for den flerspråklige søkemotoren er at man skriver søket i et av de nordiske språkene, og programmet søker i kilder på flere språk samtidig. Søkeresultatet gis på søkespråket med henvisning til kildene, som selvfølgelig har det opprinnelige språket. Nettdordboken har som mål å gi tilgang til leksikalske, parallelle ordbokressurser på Internett.

## Hva brukes språkteknologi til?

### - kommunikasjon mellom menneske og maskin

Språkteknologi forenkler og forbedrer kommunikasjonen mellom mennesker og maskiner, og den hjelper mennesker til å kommunisere med hverandre. **Målet for språkteknologien er å kunne kommunisere med maskinene ved hjelp av naturlig tale.** Der er vi ikke ennå, men det er det man strekker seg etter. I hverdagen kan språkteknologien være et nyttig hjelpemiddel ved å effektivisere arbeidsprosesser. Nesten all informasjon behandles, lagres og finnes fram av datamaskiner i vårt samfunn i dag, og mye av den skriftlige kommunikasjonen foregår via datamaskiner.

### - språkkontrollverktøy

De fleste som bruker pc, kjenner til at tekstbehandlingsprogrammene har stave- og en viss grad av grammatikkontroll. At dette er resultatet av å kombinere og omgjøre til praktisk bruk kunnskap om rettskriving, grammatikk og teknologi, er de færreste klar over. For at disse programmene skal bli bedre, trengs mer kunnskap om språket og hvordan det brukes. En del av den kunnskapen kan man få ved å undersøke hvordan ord brukes, hvordan setninger konstrueres i store tekstmengder, og så lage statistiske beregninger på hvor ofte de ulike konstruksjonene forekommer. Gjennom avdekking av regelmessigheter i de språklige byggsteinene kan man skrive regler (dataprogrammer) som datamaskinen kan anvende til kontroll av nye tekster.

### - talegjenkjenning

Et annet bruksområde for språkteknologi er talegjenkjenning i interaktive tjenester, som for eksempel automatisk billettbestilling på telefon. Kunden som ringer, blir spurt om hva han/hun vil ha, svarer og får opplyst et referansenummer og hvor billettene kan hentes. Hele transaksjonen er utført gjennom kundens kommunikasjon med en datamaskin. Talegjenkjenning brukes også på sykehusene for å spare tid og personale. Legen analyserer røntgenbilder av en pasient og leser inn sine funn i pasientens elektroniske journal. Mellom legen og journalen har datamaskinen en talegjenkjenner som "oversetter" legens talestrøm til tekst i et dokument (journalen). Legen kan selv kontrollere på skjermen foran seg at teksten blir korrekt. Man har da

# SPRÅKRÅDET

spart tid for en fagperson som må skrive etter legens diktat. Og man har spart tid for alle involverte ved at journalen er klar for neste ledd i sykehusbehandlingen umiddelbart etter at røntgenlegen har lagret journalen. Det er ikke vanskelig for noen å forestille seg hvor viktig dette kan være i en krisesituasjon der hvert sekund teller.

Forutsetningen for at talegjenkjenning skal fungere, er at man under utviklingen har tilgang til store mengder bearbeidet og godt merket lyd i form av opptak av naturlig tale. For norsk er det viktig at lydopptakene har god spredning på antall dialekter i og med den utbredte aksepten for å bruke dialektene i alle sammenhenger. Vi har heller ikke noen standarduttale for norsk muntlig. Avvik i uttalen fra personer med andre morsmål enn norsk må også være representert. Tilrettelagt og godt merket fagterminologi er en annen viktig ressurs som datamaskinen må ha tilgjengelig. Den brukes for å kontrollere at talegjenkjenneren tolker den inngående lydstrømmen som gjelder fagordene, riktig.

## *- kunstig tale*

Kunstig eller syntetisk tale er et annet bruksområde som ikke er vanskelig å se nytten av: Brukeren av en digital tjeneste kan velge å få lest opp innholdet i stedet for å lese selv. Departementene har hatt en slik tjeneste koplet opp til sine nettsider, men tidligere i år ble den vurdert som for kostbar. Etter protester fra brukerne har beslutningen blitt omgjort. Grunnen til at tjenesten er dyr, er selvsagt at språkteknologi koster. Det er dyrt å utvikle disse produktene. Og blir det svært dyrt om hvert enkelt firma selv må samle og bearbeide alle nødvendige språklige underlag fra grunnen av. Regjeringen innså selv at konsekvensene av å ta bort tjenesten var uheldig. Brukere med dårlige leseferdigheter eller nedsatt syn ville ikke lenger ha samme tilgang til offentlig informasjon. Det ville også fått konsekvenser for videre utvikling av kunstige stemmer: Signaleffekten ville bli at dette er ikke noe det offentlige vil satse på, og følgelig kan man heller ikke regne med at det offentlige vil bidra med midler til videreutvikling. Forbedring av kunstige stemmer forutsetter at de tas i bruk. Tar mange dem i bruk, vil det spore til videre utvikling, og videre utvikling fordrer tilgang til godt merket akustisk materiale (opptak av naturlig tale).

## *- informasjonssøking*

Kunnskap om et språks semantikk bidrar i dag til at vi har mer effektive informasjonssøkesystem enn vi hadde bare for noen få år siden. Og utviklingen skjer hurtig på dette området. I avanserte søkemaskiner kombineres kunnskap om språk, grammatikk og semantikk med avansert sannsynlighetsberegning. Resultatet er stadig raskere sortering i informasjonsmengdene. Men det er et stykke fram før vi kan bruke stemmen til fortelle datamaskinen hva vi leter etter. Avanserte amerikanske mobiltelefoner kan i dag opereres ved hjelp av stemmen.

## *- hjelpemidler*

Små programmer basert på språkteknologisk kunnskap og kompetanse legges i dag inn i ulike dagligdagse gjenstander: vi har fått GPS i bilene der en stemme forteller hvor, og hvordan, du skal kjøre for å komme raskt fra A til B (norske kart!).

Hus og ulike hjelpemidler for handikappede kan i dag styres med stemmen, men dette er dyre løsninger og det er et stykke fram før det kan bli allemannseie. Den dagen det blir det, øker også sikkerheten f.eks. i forhold til innbrudd i hus: Huset

# SPRÅKRÅDET

godtar ikke andre stemmer enn dem som er lagt inn, og alarmen går når andre trenger seg inn uten aksept.

## Konklusjon

En god start for en norsk språkbank er å sørge for å ta vare på det omfattende arbeidet som ble gjort av NST, og la andre nyte godt av det ved å løse ut de akustiske basene og gjøre dem tilgjengelige. Skal vi ha en norsk språkpolitikk, må vi ha en norsk språkteknologi og en norsk språkbank.

Uansett hvilken vei vi velger for norsk språkpolitikk, kommer vi ikke utenom språkteknologi. Satser vi på norsk språkteknologi, kan språkteknologiske produkt brukes i en innovativ og framtidsrettet satsing på kunnskapssamfunnet. Digitalisering og tilgjengeliggjøring av vår kulturarv vil være enklere med norsk språkteknologi. Velger vi å satse på en politikk med parallellspråklighet, kan norske språkteknologiske produkter brukes til å styrke valget og gjøre det enklere for nordmenn å håndtere både norsk og engelsk.

Velger man å la være å satse på norsk språkteknologi, vil det selvfølgelig fortsatt finnes informasjon tilgjengelig på norsk, men de store mengdene, de avanserte verktøyene og produktene vil foreligge på engelsk. Det vil medføre at de unge som kan engelsk godt nok, tar i bruk teknologien, mens eldre og mange andre uten den nødvendige kunnskapen i engelsk stenges ute. Og Norge som et høyt utdannet kunnskapssamfunn vil sakke akterut. Trusselen om at norsk språk er utrydningstruet, kan bli en realitet raskere enn vi i dag kan forestille oss, og dette fordi norsk stadig blir marginalisert og nedprioritert.