Språkteknologi på norsk

Av Jon Grepstad

- Dei språk som ikkje blir digitaliserte, kjem til å døy ut som fullverdige språk i moderne samfunn.
- Dersom nynorsken ikkje blir teken vare på i den digitale verda, døyr nynorsken ut.
- Dersom vi ikkje tematiserer den kjensgjerninga at vi har to målformer her i landet, kjem vi til å bli sitjande att med stort sett berre bokmålsprogram.
- Styresmaktene har som mål å skape gode språkverkty og informasjonstenester både for nynorskbrukarar og bokmålsbrukarar.
- Da må styresmaktene òg setje i verk tiltak som sikrar at ein når dette målet.

1 Informasjonsteknologi og språk

"Computing is not about computers any more. It is about living," seier Nicholas Negroponte i boka Being digital frå 1995. Med denne utsegna minner han oss om at informasjonsteknologien grip stadig djupare inn i kultur, samfunnsliv, livsmønster og sosiale omgangsformer. Ja, jamvel tilhøvet vårt til tid og rom er i ferd med å bli endra. Kor avhengige vi er av denne teknologien, og kor sårbare vi er, viser 2000 års-problemet.

Også når det gjeld språk og språklege tilhøve, får informasjonsteknologien stadig meir å seie. På nettet ligg mange språksamfunn berre nokre tastetrykk unna. Vi er på veg inn i "den globale landsbyen", som Marshall McLuhan tok til å tale om alt i 1962. For framandspråkopplæringa er denne nærleiken eit stort gode. For mange av oss er han ein rikdom i fritida. Men dette virtuelle eller digitale grannelaget gjer òg at presset frå engelsk aukar i samfunn der tiltrua til eige språk og eigen kultur ikkje er rotfest nok.

For språk som har fått utvikla teiknsett som kan nyttast på datamaskinar, har informasjonsteknologien gjort formidling og reproduksjon av tekstar svært billig. Det kan små språksamfunn tene på. Men mange dataprogram finst berre på dei store språka. Det kan medverke til ei digital språktyning.

I Noreg finst det knapt brukarprogram i fullstendige nynorskversjonar, og berre 10-15 prosent av dei pedagogiske programma i skulen er på nynorsk - trass i vakre formuleringar i KUFs IT-plan for 1996-99, og trass i at Språkrådet alt i 1994 peika på overfor KUF at det burde vere full språkleg likestilling i slike program. Somme av programma er attpåtil på dårleg nynorsk. Berre 4-5 prosent av norskspråklege dokument på Internett er på nynorsk. Og departementa sin informasjonstenar ODIN, som etter mållova har plikt til å ha minst 25 prosent nynorsk, har ikkje meir enn rundt 10 prosent av dokumenta i nynorsk målform.

I denne artikkelen skal vi sjå nærmare på språkteknologi og norsk språk, med særleg vekt på nynorsk.

2 Kva er språkteknologi?

"Språkteknologi" kan kort definerast som "dataprogram som behandlar naturleg språk" eller kanskje rettare: "språkmodular i dataprogram som behandlar naturleg språk." Med "dataprogram" tenkjer vi da ikkje berre på program som ein køyrer på ein datamaskin, men også program innebygde i moderne tekniske innretningar, t.d. program som gjer det mogleg å styre maskinar med røysta.

Eit hovudpoeng i denne artikkelen er at språkteknologi vil spele ei stadig viktigare rolle i tida som kjem, og at språk som vil leve vidare som fullverdige bruksspråk på alle felt i eit samfunn, må takast vare på i språkteknologien. I dag gjeld dei mest avanserte programma engelsk - og somme andre store språk.

Døme på språkteknologi er:

(a) korrekturprogram og anna skrivestøtte
(b) informasjonssøking
(c) maskinomsetjing
(d) taleteknologi (talegjenkjenning og talesyntese)

Korrekturprogram kjenner mange av oss som retteprogram i tekstbehandling. Programma er blitt meir og meir avanserte, og dei blir tettare integrerte i skriveprogramma. Vi har ikkje berre gammaldags stavekontroll, men også synonymordlister, for somme språk grammatikkontroll, og kanskje også stilkontroll. Avstanden til skrivemaskinen har auka kraftig.

Dei nyaste tekstbehandlingsprogramma omfattar også former for informasjonsbehandling. Word 97 har "autosamandrag" - ein kan be programmet samanfatte eit dokument, velje mellom fire typar oppsummeringar, og avgjere kor omfattande dei skal vere. Men språket må førebels vere engelsk, og resultatet kan nok vere så som så.

Informasjonssøking er kjend mellom anna frå Alta Vista på nettet. Ein kan søkje på stikkord, og ein kan avgrensa søka til bestemte språk. Programmet som ligg bak, er altså i stand til å identifisere språket i tekstane. Går vi inn på departementa sin informasjonstenar ODIN, eller på nettsidene til Statistisk sentralbyrå, kan vi søkje på liknande vis, men vi må taste inn søkjeorda på begge målformer -- fordi vi på norsk vantar ei synonymordliste eller ein tesaurus som går på tvers av bokmål-nynorsk. Og søkjeprogramma byggjer framleis på indekserte ord, ikkje på emneord eller omgrep.

Maskinomsetjing spelar ei stadig viktigare rolle i internasjonal kommunikasjon og i fleirspråklege samfunn. Omsetjingssystema blir nytta av internasjonale organisasjonar (mellom anna i EU-systemet), i forretningslivet og i informasjonstenester. I Noreg har vi fått omsetjingsprogrammet Nyno, som omset frå bokmål til nynorsk, og som stundom blir nytta til råomsetjing av lærebøker.

Di klarare avgrensa saksområdet og vokabularet er, di betre blir omsetjingane. Systemet Taum Meteo i Kanada har i 20 år kvar time 24 timar i døgeret omsett automatisk vermeldingar frå engelsk til fransk, frå 1989 også andre vegen. I løpet av eitt døger omset Taum Meteo 40 000 ord, og berre 5 prosent treng rettingar. Program som skal dekkje allmennspråk, krev langt meir føre- og etterarbeid om omsetjingane skal bli gode.

Omsetjingstenesta på Alta Vista, som det leiande maskinsomsetjingsfirmaet Systran står for, er eit døme på det. Men tenesta kan likevel vere nyttig. Her er ei Alta Vista-omsetjing frå portugisisk til engelsk av ein omtale av "Web Translator", eit innpluggingsprogram som kan omsetje nettsider automatisk mens ein surfar:

"The Power Pro Translator and the Web Translator allow that internauta gets the text in Portuguese without needing to abandon the interface of browser. It is enough to clicar in a button in the bar of tools of the navigation program and the translation comes in few seconds. Clearly that the translation time depends on the size of the text."

Taleteknologi - talegjenkjenning og talesyntese - er kanskje det feltet som har fått størst merksemd den siste tida. Det har skjedd mykje sidan IBM presenterte Shoebox-maskinen som kunne lese tal på Verdsutstillinga i New York i 1964. Særleg dei siste åra har det vore gjort store framsteg. Om ti år kan det vere like naturleg å snakke til ein datamaskin som å nytte tastatur eller mus, spår somme språkteknologar. Eller for å sitere Bill Gates: "Speech is the future of computing itself." Talekontroll vil kunne bli grensesnitt til mange moderne tekniske innretningar, blir det hevda.

Datamaskinar som kan styrast med røysta, finst allereie. Nyleg har det kome i handelen dikteringsprogram frå Dragon Systems, IBM og Lernout & Hauspie som let brukaren snakke inn tekst til datamaskinen. Etter som ein dikterer, blir teksten vist i tekstbehandlingsprogrammet, reknearket eller databaseprogrammet. Før ein tek programmet i bruk, trenar ein det opp til å bli kjent med eins eiga røyst og uttale. Programma behandlar samanhengande tale - det er ikkje lenger naudsynt å diktere med pause mellom kvart ord. Ein kan redigere teksten med røysta, setje margar, velje font, be om kursiv eller halvfeit. Somme av programma kan kjøpast med ordlister tilpassa medisin, jus, journalistikk eller andre fag. I dei mest avanserte programma kan ein sjølv byggje opp vokabular for ulike emne. Dragon Naturally Speaking, til dømes, kan gå gjennom filene på disken, sjå på ordfrekvens og skrivemåte, og generere ei ordliste tilpassa dei behova brukaren har. Dikteringsprogramma gjeld førebels berre engelsk og somme andre store språk. Og prisane har falle mykje dei siste åra. Enklare versjonar kostar eit par tusen kroner eller mindre. Dei mest avanserte versjonane kostar monaleg meir. Ein god mikrofon og lite bakgrunnsstøy er ein føresetnad for gode resultat.

Taleteknologien kan brukast på mange område. Telefontenester er eitt felt. Ein tastar seg fram til den informasjonen ein skal ha. I Noreg lanserte Telenor i fjor ein e-postlesar som les e-posten høgt. Ein ringjer eit nummer og får lese opp kor mange meldingar som ventar i postkassa, kven som har sendt dei, og kva dei handlar om - og ein kan velje ut dei meldingane ein vil programmet skal lese. Etterpå kan ein velje om ein vil slette meldingane eller ikkje.

Dei som vil ha ein forsmak av kva som kanskje ventar oss, kan ringje telefonnummer 63 84 85 22, der ein møter ein demoversjon av Telenor FoUs automatiske nummeropplysning, og spørje etter telefonnummeret til nokre av dei tilsette i avdelinga, t.d. "Natvik", "Tore" eller "Einar Flydal".

Eller ein kan gå ut på nettet til denne adressa - http://www.pcww.com/index.html - og gratis hente eit "klippe-og-snakke-program", installere det på mindre enn to minutt, starte nettlesaren og be programmet lese teksten på nettsidene høgt for seg. Eller hente eit dokument inn i tekstbehandlingsprogrammet og få det lese opp. Men - teksten må vere på engelsk.

3 Utfordringar for norsk - bokmål og nynorsk

Ein del av språkteknologien har teke utgangspunkt i behova til funksjonshemma eller spesielle yrkesgrupper. Typisk i dag er det at teknologien blir allmenngjord. Programma blir utvikla for stadig nye felt i daglegliv og yrkesliv. Stundom kan nok utsiktene vi blir presenterte for, minne oss om hovudpersonen i Heinrich Bölls novelle "Es wird etwas geschehen" - mannen som var så effektiv at han ikkje visste kva han dreiv med. Men teknologien kan også dekkje reelle behov.

Den mest avanserte språkteknologien finst for engelsk og somme andre store språk. Det ligg mykje utviklingsarbeid bak språkmodulane i programma. Og utviklingsarbeidet kostar pengar. Eit dikteringsprojekt for norske sjukehus i regi av Philips vart for nokre år sidan stogga pga. pengemangel, skreiv Computerworld Norge 27. februar i fjor. I same nummeret sa Bjørn Tore Hoem i IBM at IBM sitt dikteringsprogram ViaVoice - eitt av dei leiande programma i dag - ikkje vil kome på norsk. Marknaden er for liten, vi har to skriftspråk å ta omsyn til, og mange dialektar. Utviklingsarbeidet kostar for mykje, meinte Hoem.

Interessant er det da å sjå at det nyskipa firmaet Nordisk Språkteknologi AS på Voss, i samarbeid med det belgiske taleteknologifirmaet Lernout & Hauspie (eit av dei leiande språkteknologifirma i verda, der Microsoft kjøpte seg inn hausten 1997), no er i ferd med å utvikle taleteknologiprogram for norsk, svensk og dansk. I utviklingsarbeidet førebur NST utvikling for nynorsk, men nynorsk ligg enno ikkje i dei konkrete framdriftsplanane, opplyste firmaet i på ein konferanse i januar i år. "Vi er i ein situasjon som liknar på situasjonen til forlag som lagar lærebøker til skulemarknaden: Utan ekstern finansiering er det ikkje råd å driva nynorskproduksjon med lønsemd," hevda firmaet - og fekk straks svar frå representanten for det finske firmaet Lingsoft, som nyleg har gjort gode pengar på å selje avanserte korrekturprogram for Microsoft Office 2000 både på bokmål og nynorsk.

Norsk språk, bokmål og nynorsk, er på fleire felt i samfunnet utsett for sterkt press frå engelsk. Språksosiologar snakkar om faren for "domenetap" - at norsk på bestemte område blir erstatta av engelsk. Det kan gjelde høgteknologiske område og delar av forretningslivet. For nynorskbrukarar er domenetap i høve til bokmål velkjent. Mange konvensjonelle dataprogram tek ikkje omsyn til at vi har to skriftspråk. Dersom vi ikkje tematiserer den kjensgjerninga at vi har to målformer her i landet, blir vi sitjande att stort sett berre med bokmålsprogram. Det gjeld også skuleverket.

Styresmaktene meiner språkteknologien kan auke det engelske presset på norsk dersom det ikkje blir teke særskilde åtgjerder. I St.meld. nr. 13 (1997-98), som Kulturdepartementet la fram i desember 1997, vart spørsmåla drøfta, og det vart gjort framlegg om at det blir skipa eit sekretariat for språkteknologi i tilknyting til Norsk språkråd. Meldinga seier mellom anna:

"Kompleksiteten i dei nye språkteknologiske produkta gjer at ein ikkje ved enkle grep kan tilpasse produkta til bruk i Noreg. Skal det kunne skje, må vi ha tilgjengeleg i nynorsk- og bokmålsversjon dei språkmodulane som ligg føre for engelsk i den originale utgåva. Dersom vi ikkje er i stand til å gjere dette tilretteleggingsarbeidet, som igjen må vere grunnlagt på norskspråkleg forskings- og utviklingsarbeid, vil mange brukarar i Noreg truleg gå over til å nytte engelskspråklege versjonar for å få del i det sterkt auka brukspotensialet i dei nye produkta. Slik vil det mellom anna kunne gå innanfor konkurranseutsett eksportretta næringsliv. Dette vil truleg smitte over til undervisning og forsking. Dei samla språkpolitiske verknadene av dette kan bli store. Vi må derfor ruste oss til kamp for i det heile å kunne nytte norsk på alle bruksområde i eit stadig meir teknologisk samfunn."

Til det siste har leiaren for nynorskseksjonen i Norsk språkråd, professor Kåre Lilleholt, skrive i ein artikkel i Språknytt 4/1998:
"Her ottast eg meir for jamstillinga mellom bokmål og nynorsk. Dersom vi får nye språkverkty i bokmålsversjon, men ikkje i nynorskversjon, må vi rekne med at det kan gå ut over nynorskbruken. Det skal noko til før vi skiftar frå norsk til engelsk, men så godt som alle nordmenn kan bruke bokmål like godt som nynorsk. Da kan det nye hjelpemiddelet vere det som skal til for at bokmålet tek over."

Det trur eg er eit svært viktig tilskot til analysen.

Og da kan det vere god grunn til å nemne at EUs Euromap-rapport frå september 1998, ein rapport som skulle kartleggje situasjonen for språkteknologien i EU- og EØS-området, ikkje nemner tospråkssituasjonen med eitt ord i den delen av rapporten som gjeld Noreg, trass i at den norske rapporten er laga i Noreg av firmaet IDE AS.

4 Grunnlaget må vi sjølve leggje

Grunnlaget for språkteknologien er "språkressursar", store samlingar med tekst og tale, omfattande elektroniske ordlister og synonymordlister koda på bestemte måtar. Det trengst både allmennspråk og fagspråk. Talematerialet må omfatte personar av begge kjønn og i ulike aldrar, med ulik uttale og i forskjellige lydmiljø.

Representantar for fagmiljøa har ved fleire høve peika på at arbeidet med gjenbrukbare språkressursar må få særskild støtte frå staten. Felles ordlister, talemateriale og andre basisdata må tilretteleggjast og gjerast tilgjengelege slik at dei kan nyttast av firma som utviklar og produserer språkteknologi.

Ei slik investering vil vere eit viktig tiltak for å verne om norsk språk i tida som kjem. Og det vil også vere ei ordning som gjer det enklare å ta vare på både nynorsk og bokmål i språkteknologien.

I Noreg blir det i dag arbeidd med språkteknologi først og fremst ved universiteta, SINTEF, Telenor FoU og Nordisk Språkteknologi AS, og dessutan ved somme forlag. Frå Finland leverer Lingsoft, eit av dei leiande språkteknologifirma i verda, stavekontroll og orddelingsprogram for bokmål og nynorsk til Microsoft Office 2000. Programma skal vere monaleg betre enn dei som ligg i Word 97. Nynorskversjonen byggjer på Nynorskordboka, som Norsk språkråd har vore med på å utarbeide. I tillegg finst m.a. EU-prosjektet "Scarrie", som skal leggje grunnlaget for automatisk korrekturlesing for skandinaviske språk. For norsk omfattar prosjektet berre bokmål pga. knappe løyvingar. Frå før finst korrekturprogrammet Tansa, som blir nytta til korrekturlesing av m.a. VG, Dagsavisen, Bergens Tidende, Nynorsk Pressekontor og andre. Omsetjingsprogrammet Nyno, som er nemnt tidlegare, kom i fjor haust i versjon 3.0. Versjon 3.1 er like rundt hjørnet.

Mellom fagmiljøa ved universiteta i Oslo, Trondheim og Bergen vart det for eit par år sidan skipa eit eige nettverk, Nasjonal infrastruktur for språkteknologi (NIFST). Nettverket fekk fastare form hausten 1998. Det er blitt utvikla eit program for automatisk grammatikk-koding av tekst på bokmål og nynorsk (ein "morfosyntaktisk taggar"), maskinleselege ordbøker for bokmål og nynorsk ("NorKompLeks"), og det blir arbeidd med å byggje opp korpora.

5 Mål og middel

I St.meld. nr. 13 (1997-98) formulerer regjeringa denne målsetjinga når det gjeld informasjonsteknologi og norsk språk:

"Målet er å kunne gi både nynorsk- og bokmålsbrukarane gode reiskapar til støtte for språkbehandling og å tilby relevante informasjonstenester o a på begge målformer."

Spørsmålet er da kva tiltak styresmaktene set i verk for å sikre at ein når dette målet. I stortingsmeldinga peikar Kulturdepartementet på at det krevst tiltak på fleire felt: "Det trengst eit offensivt handlingsprogram der både juridiske, administrative, språkteknologiske og økonomiske verkemiddel blir tekne i bruk."

Førebels har styresmaktene teke initiativ til å få skipa eit sekretariat for språkteknikk i tilknyting til Norsk språkråd. Språkrådet har søkt om tre stillingar for neste års budsjett, og det blir no sett i gang utgreiing av hovudoppgåver og kompetansekrav i samband med stillingane. Det er såleis nokså ope kva som blir arbeidsoppgåvene, men målsetjinga er å fremje språkteknologi på norsk, både bokmål og nynorsk. Å tale om ein strategi frå styresmaktene si side er for tidleg.

Vidare er språkteknologi peika ut som eit viktig felt i regjeringa sin IT-plan for 1998-2000, "Norge - en utkant i forkant". I oppfølginga av planen har Nærings- og handelsdepartementet, som har samordningsamsvaret for regjeringa sin IT-politikk, bedd Noregs forskingsråd vurdere "hvordan utviklingen av norsk språkteknologi kan styrkes". Forskingsrådet har på si side nyleg teke kontakt med Kulturdepartementet for å drøfte ymse spørsmål i samband med oppbygging av språkressursar. Språkteknologi er også eit sentralt tema i regjeringa sin handlingsplan for funksjonshemma, som vart lagd fram i St.meld. nr. 8 (1998-99).

Noregs forskingsråd har løyvd midlar til eit forprosjekt for eit nasjonalt korpus. Forprosjektet skal førebu ein omfattande søknad om etablering av eit nasjonalt korpus for språkteknologi. Forprosjektet er eit samarbeid mellom eit konsortium som består av NTNU, Universitetet i Bergen, Universitetet i Oslo, Nordisk Språkteknologi, SINTEF og Telenor FoU. Prosjektleiar er professor Torbjørn Svendsen ved NTNU. I februar i år vart det halde eit seminar i Trondheim for å drøfte det vidare arbeidet med eit nasjonalt korpus. Møtet samla fagmiljøa i Noreg, og hadde òg ein representant for British National Corpus. På bakgrunn av møtet blir det no førebudd ein meir omfattande søknad til Forskingsrådet. Arbeidet vil omfatte både nynorsk og bokmål.

I tildelingsbrevet for 1999, dvs. det brevet som gir den økonomiske løyvinga, blir Noregs forskingsråd og Statens nærings- og distriktsutbyggingsfond elles pålagde å spele ei aktiv rolle for å stimulere til utvikling og kommersialisering av språkteknologiske produkt og tenester.

Men når det gjeld tiltak som sikrar at nynorsk blir teke vare på i språkteknologien, kviler ansvaret på Kulturdepartementet. Departementet har formulert gode mål i St.meld. nr. 13 (1997-98). No er tida komen til å finne fram til tiltak som kan sikre at måla blir nådde.

Nokre tilvisingar:

- Informasjonsteknologi og norsk språk. Eit oversyn over kva Storting og departement har sagt om emnet. Oslo 1999. http://home.sol.no/~gjon/artiklar.htm

- Språknytt nr. 4/1998. (Temanummer om språkteknologi). Oslo: Norsk språkråd. http://www.sprakrad.no/snytt984.htm
- Statssekretær Odd Hellesnes: Regjeringa sin strategi for IT-utvikling. 19. januar 1999. http://odin.dep.no/nhd/ataler/1999/oh990119.html
- Referat frå konferansen "Språkteknologi på norsk", Oslo 12.-13.10.98: http://www.hit.uib.no/spraaktek98/referat.html
- Referat frå arbeidsmøte om planlegging av eit nasjonalt korpus, Stjørdal 12.2.99. http://pan.hf.ntnu.no/nifst/nasjonalt_korpus/korpusseminar.html
- Omtale og vurdering av omsetjingsprogrammet Nyno: http://www.hf.uio.no/tekstlab/bulletin/bull1_98/nyno.html
- The Euromap Report: http://www.linglink.lu
- Dragon Systems, Inc.: http://www.dragonsys.com/
- IBM ViaVoice: http://www.software.ibm.com/speech/
- Lernout & Hauspie: http://www.lhs.com/
- Lingsoft, Inc.: http://www.lingsoft.fi/
- Nordisk Språkteknologi AS: http://www.ide-as.com/nst/
- Nynodata AS: http://www.nynodata.no/
- Systran: http://www.systransoft.com/

[Mål og makt 1999]