

# Norsk språkbank – ein framtidig ressurs for transløtarar?

Torbjörg Breivik

Språkrådet, Noreg

## Summary

The primary argument in this paper is that terminology will be an important part of a human language resource collection for a specific language. Translators need terminology and machine translation software needs bilingual terminology resources as input. Traditionally translators have their own memory sources and terminology lists built through years. Translators working in translation companies may have access to large and validated terminology resources, but in Norway there are just a few of those. In Norway we plan to build a public human language resource collection to promote language technology research and innovation. An important and fundamental part of such a database will be reliable, standardised and bilingual terminology resources. These resources will be useful for translators as they will contain high quality and recommended terminology for various professions and subjects. In this paper I give a short status for the project and then discuss how terminology and translation resources can fit into the project. In the end translators are given a challenge: their knowledge is of great value to the project and their contribution will be most appreciated. We will all gain a lot through cooperation.

Når ein omset tekst frå eitt språk til eit anna, er det vanleg at kvar omsetjar held seg med personlege omsetjarminne i programmet han eller ho nyttar i arbeidet. I praksis må ein slik arbeidsmetode kallast maskinstøtta omsetjing. Vi har enno langt att før vi kan la maskinane omsetja for oss. I mange år framover vil transløtøren vera ein viktig aktør i arbeidet med omsetjing av tekstar frå eitt språk til eit anna. Det betyr ikkje at vi ikkje skal gjera programma betre og meir tenlege for desse brukarane. Ein framtidig norsk språkbank kan vera til hjelp for transløtarar ved at han inneheld råvarer for dei som arbeider med omsetjarprogram. Råvarene vil vera tospråklege terminologibasar og tospråklege korpusar. Det langsiktige målet for språkbanken er å få til ei velfungerande og brei samling norske språkressursar av høg kvalitet der innhaldet følgjer internasjonale standardar. Språkressursane skal vera tilgjengelege for forskning og for kommersielle aktørar i næringslivet.

Artikkelen startar med å gje bakgrunn og status for kvar vi er i dag når det gjeld å etablera ein språkbank i Noreg, og ser deretter nærare på kva delar av innhaldet i ein språkbank som kan vera til nytte for transløtarar i deira arbeid. Det mest aktuelle innhaldet vil vera terminologiressursar. Men eg vil også gje transløtarane ei utfordring: Dei kan medverka til oppbygging av språkbanken ved å gje banken terminologisamlingar. Gjennomarbeidde og kvalitetssikra terminologisamlingar (databasar) vil vera ein svært nyttig og nødvendig del av ein språkteknologisk infrastruktur for norsk.

## Bakgrunn

Nærings- og handelsdepartementet laga i 2000 e-Norge-planen<sup>1</sup>, der regjeringa sette seg som mål å skapa eit informasjons- og kunnskapssamfunn for alle. Planen hadde konkrete bestillingar til andre departement som galdt oppgåver og arbeid som måtte gjerast på ulike område for å realisera det samfunnet regjeringa såg for seg. Til Kultur- og

---

<sup>1</sup> e-Norge, versjon 1, NHD juni 2000

kyrkjedepartementet (som Språkrådet er underlagt) kom det ei bestilling på ein strategisk politisk plan for norsk språk og IKT, der det vart spurt etter konkrete tiltak for å ta vare på og utvikla det norske språket i eit moderne informasjons- og kommunikasjonssamfunn.

Språkrådet svara med å levera ein handlingsplan for norsk språk og IKT<sup>2</sup> til Kultur- og kyrkjedepartementet i 2001. Det største tiltaket i planen var å samla språklege ressursar som tekst, tale, ordbøker, termbasar og verktøy, tilordna dei og gjera dei tilgjengelege for forskning og næringsliv. I 2002 blei Språkrådet bede om å utarbeide ein plan for etablering av ei slik samling. Planen blei skriven og lagt fram for Kultur- og kyrkjedepartementet og Nærings- og handelsdepartementet (som begge var oppdragsgjevarar) i oktober 2002<sup>3</sup>. Til dagleg blir samlinga omtalt som norsk språkbank.

Forskningsrådet hadde lenge hatt i gang arbeid på informasjonsteknologiområdet, og i 2002 blei det sett i gang eit eige program for språkteknologi, kalla KUNSTI<sup>4</sup>. Det var ein føresetnad for programmet at ein norsk språkbank fanst eller i alle fall skulle byggjast opp parallelt med dette programmet.

Stortingsmeldinga ”Norsk kulturpolitikk fram mot 2014”<sup>5</sup> har eit langt kapittel om kor viktig norsk språkteknologi og ein norsk språkbank er for å ta vare på det norske språket i teknologien. Der blir det særleg streka under at om alle i Noreg skal få tilgang til dei nye tenestene og dei nye produkta som no blir utvikla, må dei bli tilgjengelege på norsk. Språkbanken skal etablerast for å minska kostnadane med utvikling av norskspråklege tenester og produkt. Konklusjonen er overraskande nok ikkje at vi må setja i gang arbeidet med den norske språkbanken snarast råd er, men at det ”På kort sikt er det difor ikkje aktuelt å setja av statlege midlar til å utvikla ein norsk språkbank” (s. 197).

### **Kvar står vi i dag?**

Nordisk Språkteknologi Holding AS var eit norsk firma som utvikla språkteknologiske produkt for dei skandinaviske språka (norsk, svensk og dansk). Firmaet gjekk konkurs i november 2003. Språkrådet hadde tett kontakt med buet om å få overta dei språkressursane vi visste var verdfulle for andre. I 2006 blei buet kjøpt av eit konsortium: universiteta i Bergen, Oslo og Trondheim, IBM Norge og Språkrådet. Målet med kjøpet var å sikra at ikkje verdfullt arbeid gjekk tapt for ettertida, og særleg då med tanke på ein norsk språkbank. Innhaldet i buet måtte sorterast, sikrast, lagrast og leggjast over i nyare programvare, eit oppdrag som vart sett ut til Aksis ved Universitetet i Bergen. Oppdraget danna grunnlag for å søkja om midlar frå Forskningsrådet, og Forskningsrådet løyvde 1 030 mill. kr. til prosjektet. Kjøpet av buet er viktig av fleire grunnar: her er mykje godt arbeid gjort, mange har bruk for denne typen språkressursar, og med tanke på ein norsk språkbank vil dette materialet spara oss for mykje arbeid og utgifter.

Eigarane har danna eit interimsstyre for Norsk språkbank med ein representant frå kvar. Språkrådet har styreleiar og er administrativt ansvarleg for å driva arbeidet vidare inntil det

---

<sup>2</sup> Handlingsplan for norsk språk og IKT, Språkrådet 2001

<sup>3</sup> Samling og tilgjengeleggjering av norske språkteknologiressursar, Språkrådet 2002

<sup>4</sup> Kunnskapsutvikling for norsk språkteknologi, Noregs forskingsråd 2001

<sup>5</sup> St.meld. nr 48 (2002-2003) Kulturpolitikk fram mot 2014

ligg føre eit politisk vedtak om å oppretta norsk språkbank<sup>6</sup>. Når språkbanken formelt er oppretta, har interimsstyret fullført si oppgåve og innhaldet vil inngå i den nye organisasjonen. Målet med kjøpet var samanfallande med målet for etablering av ein språkbank: å få til ei velfungerande og brei samling norske språkressursar av høg kvalitet der merking og tilretteleggjing for gjenbruk følgjer internasjonale standardar. Språkressursane skal vera tilgjengelege for forskning og for kommersielle aktørar i næringslivet.

I 2008 vil interimsstyret konsentrera seg om å gjera ressursane klar til distribusjon. Avtalar for lisensiering av ressursane er under arbeid, og dei er utarbeidde med sikte på at dei skal kunna brukast også av den endelege språkbanken.

### **Innhald i språkbanken**

I dag har interimsstyret hand om språkressursane frå konkursbuet etter Nordisk Språkteknologi Holding AS på Voss. Det mest verdfulle er, som før nemnt, opptaka av tale (dei akustiske basane) med tilhøyrande leksikon. Det er akustiske ressursar for norsk, svensk og dansk. Dei er systematisk og godt merkte, og dei følgjer standardane som var tilgjengelege då dei blei laga (1999-2000). Det er i 2008 ikkje sett av ressursar til å oppdatere merkinga til dagens standardar på området.

Interimsstyret vil arbeide for at styresmaktene skal gjera vedtaket om å oppretta ein språkbank så snart som mogleg, og vi håper det kan setjast av midlar i neste statsbudsjett (for 2009). Det er varsla to språkmeldingar i 2008, og i den som gjeld språkpolitikken framover, reknar vi med at språkteknologi får stor plass, og at språkbanken er ein naturleg del av satsinga på det området.

Når arbeidet med den endelege språkbanken startar, må planen frå 2002 reviderast, ein må gjera prioriteringar for innsamling ut frå situasjonen på det tidspunktet, og ein må oppdatere oversynet over kva som finst hos ulike aktørar. Når det gjeld innhaldet i ein framtidig språkbank, vil ein følgja dei internasjonale tilrådingane og standardane som ligg i konseptet BLARK<sup>7</sup>. Dette konseptet blir følgt i andre land, det er utarbeidd over tid og tek høgde for at alle område av språkteknologien skal dekkjast.

### **Terminologi**

Skole, arbeidsliv, forskning – alle brukar ei eller anna utgåve av fagspråk på sitt område. Nokre gonger kan det vera vanskeleg å seia om ordet og omgrepet er ein term, andre gonger er det heilt klart at det er det. Bruk av faguttrykk gjer den fagspråklege kommunikasjonen effektiv og enkel – mellom fagfeller! Føresetnaden er at begge partar kjenner tydingsinnhaldet og har ei tilnærma sams oppfatning av det.

Det er mykje diskusjon om læring i skolen i desse tider. Det er gjort forsøk på å reisa ein diskusjon om kva plass fagtermar og fagspråk skal ha i undervisinga, utan at den debatten har nådd ut i det offentlege rommet. Pedagogar har teke til orde for at barn som ikkje kjenner og forstår omgrepa knytte til ulike fag, blir stengde ute frå reell forståing av kva desse faga

---

<sup>6</sup> Situasjonen har endra seg – sjå kommentaren til slutt i artikkelen.

<sup>7</sup> BLARK = The Basic Language Resource Kit, Stewen Krauwer /ELSNET 2003

dreier seg om. Reiskapane desse faga har utvikla for å forstå verda, vil heller ikkje vera tilgjengelege for desse barna<sup>8</sup>. Andre uttrykkjer det slik: ”Enhver kommunikatív handling mellom personer, enten det dreier seg om tale eller skrift, innebærer at man tar i bruk sjangerbestemte strukturer” (= fagspråk)<sup>9</sup>. Under desse utsegnene ligg Vygotskys (socio-interaksjonistiske) syn på barns utvikling av omgrep: barn utviklar seg sjølve, språket sitt og kva omgrep dei forstår, i samhandling med menneska (les vaksne) ikring seg. Og kva har dette med mitt tema å gjera?

Jo, vi lever i eit samfunn der dei fleste av oss blir eksponerte for faguttrykk i ulike samanhengar mange gonger om dagen. Sjølv barnehagebarn brukar i dag pc, og dei har lært eller må læra seg korleis dei skal finna fram på nettet. Og dei må læra orda som må til – ei anna sak er om dei skjønar orda. Når det gjeld søk og søkjemetodar på nettet, har det skjedd mykje dei seinare åra. Det mest interessante er utviklinga av semantisk vev (topic maps), der søkinga skjer på grunnlag av kunnskap om det semantiske innhaldet i ordet. Intelligente leksikon er eit godt grunnlag for denne søkjemetoden. Intelligente termbasar eller kunnskapsbasar som ein og kan kalla dei, er nyttig infrastruktur i ein slik samheng.

### **Terminologi i språkbanken**

Skal vi kunna søkja effektivt i databasar og på nettsider, vil vi ha stor nytte av standardiserte omgrep og faguttrykk. Språkbanken er infrastruktur for språkteknologisk forskning og produktutvikling. Ein viktig del av den infrastrukturen er ein terminologisk infrastruktur som er utvikla og tilordna for bruk i språkteknologisk samheng. Tilgang til basar med terminologi, ein- og fleirspråklege, er naudsynt for kvaliteten på mange språkteknologiske tenester og produkt. Eit eksempel er datastøtta omsetjing som translatørar arbeider med, eit anna eksempel er tale-til-tekst- system.

Ved Noregs handelshøgskole har ein i mange år arbeidd med ein kunnskapsbase for det økonomisk-administrative fagområdet, i kortform kalla KB-N-basen<sup>10</sup>. Programmet KUNSTI hos Forskningsrådet har finansiert delar av prosjektet, og ved å gjera det har Forskningsrådet også gjeve signal om at det ser denne sorten basar som ein viktig del av ei språkteknologisk satsing. Basen er mykje meir enn ei samling fagtermar ein kan slå opp i, og Marita Kristiansen omtalar han i Språknytt 2/2006 som eit språkstrategisk tiltak hos NHH. Målet med utviklinga av basen er ”i samarbeid med fagspesialistar å utvikla, kvalitetssikra og ta i bruk både den fagkompetansen og dei språkressursane som skal til for å takla skriftleg fagspråkleg kommunikasjon”. Bruken vil vera mangslungen, men Kristiansen vurderer han som eit viktig pedagogisk hjelpemiddel som vil vera til støtte i innlæringa av nye omgrep.

Frå mi side er det eit sterkt ønske at vi i Noreg kan få utarbeidd tilsvarande strukturerte, tospråklege framstillingar av omgrep og omgrepsskildringar for fleire fagområde, og at basane blir gjorde tilgjengelege gjennom språkbanken. Det vil gje heile det språkteknologiske utviklingsmiljøet i Noreg eit løft og gjera det mogleg å utvikla langt fleire tenester og produkt enn vi har i dag. Eg håpar også at språkbanken kan få høve til å formidla den programvara som er utvikla til føremålet, og såleis byggja vidare på det ein veit er bra og tenleg.

---

<sup>8</sup> Berkenkotter & Huckin, 1995

<sup>9</sup> Evensen & Løkensgard Hoel 1997

<sup>10</sup> <http://www.nhh.no/fsk/sff/kbn/>

Innhaldet i KB-N-basen, korleis han er bygd opp, og korleis han kan brukast, har vore presentert i mange samanhengar og no er han ferdig til bruk. Dette er ein type base som vil vera svært nyttig i ein språkbanksamheng. Det at eit fagmiljø brukar tid og krefter på utvikling av to (og gjerne fleir-) språklege kunnskapsbasar, styrkjer heile det norske språkteknologimiljøet. Terminologi som er samla, fagleg gjennomarbeidd, strukturert og tilrettelagd for språkteknologisk bruk, er heilt naudsynt i ein språkbank, og då inkluderer eg også enklare terminologisamlingar – dei vil også vera nyttige.

Sjukehusa har teke i bruk eit dikteringssystem på radiologiområdet der legen dikterer funna sine frå røntgenbileta direkte inn i ei tekstfil på pc-en. Legen kan kvalitetssikra fila med det same og senda ho vidare til neste behandlar etter få minuttar. Sjukehuset sparar tid og personell: den skadde kan sendast vidare raskt, og den aktuelle behandlinga iverksetjast. Og personalet kan ta seg av andre oppgåver enn å skriva ut det legen har diktert.

Dikteringssystemet er eit resultat av innsamla lyd (innlesen tekst) og tilgang til gode terminologilister. Eit system for tale-til-tekst er ikkje nok i seg sjølv – datamaskina må ha eit grunnlag å knyta lydstraumen opp mot, noko å kjenna att, og her kjem leksika og terminologi inn.

Vi har ikkje kome så langt at naturleg og ustrukturert tale kan nyttast i språkteknologiske produkt og tenester i dag, men utviklinga går i den retninga. For kvar nye elektroniske dings og duppeditt som krev språklege handlingar for å fungera, krevst det strukturerte leksika og strukturert terminologi for bruksområdet for at dei skal kunna fungera. Automatiske sentralbord og elektroniske kart i ein bil har ulike og svært avgrensa språklege underlag. Ein vaskemaskin som skal styrast med stemma, må ha enda eit anna underlag.

Ingen vil i dag lita på ein tekst som er omsett av ein datamaskin, sjølv om programvara er aldri så god. Personleg vil eg heller ikkje stø meg på hjelpa eg får frå stave- og grammatikkontrollen i tekstbehandlingsprogrammet eg brukar. Årsaka er at applikasjonane ikkje er gode nok. For lite av den kunnskapen vi har om norsk språk, er omsett til programmeringsspråka sine nullar og eittal. Nokre få program for grupper med særlege behov har kome nokre hakk lenger fordi ein har klart å omsetja fleire av reglane for norsk språk til maskinspråk. Men desse er heller ikkje så gode at du kan stø deg fullt ut på dei. Framleis må du sjølv kunna norsk rettskriving.

Hjelpesprogramma som er utvikla for grupper som dyslektikarar, har større grad av intelligens knytt til det einskilde ordet. Nye søkjesystem har også i større og større grad innebygd kunnskap om semantikk og synonymi som gjer dei betre og nyttigare enn gamle søkjemotorar som stødde seg på enkel attkjenning av ordet for å gje tilslag.

### **Generelt nyttig innhald**

I ein språkbank må det finnast ein- og fleirspråklege korpus som er tilrettelagde for bruk mellom anna i program for maskinomsetjing. Der må ein også ha verktøy for tilordning av ressursar til ein språkbank: for merking (annotering), for ekserpering, for strukturering og tilrettelegging for språkteknologisk bruk, og der skal ein ha tilgjengeleg oppdaterte norske og internasjonale standardar for format og annotering. Verktøya skal vera tilgjengelege for brukarane av språkbanken, gjerne mot at brukarane bidreg med nytt, tilordna materiale. Heilt

avgjerande er at språkbanken er staden der ein finn materiale som alltid følgjer siste versjon av standardane.

### **Terminologiarbeid i Noreg**

Knut Jonassen i Standard Norge kartla i 2005 kven i Noreg som driv med terminologiarbeid, kvar, og kva bakgrunn dei har for det – og ikkje minst: kva dei uttrykkjer av ønske på dette området. Det er ei overvekt av personar som arbeider på tekniske og naturvitskaplege fagfelt, og dei er velutdanna og har lang erfaring i yrket. Dei uttrykkjer ønske om standardiserte format, tilgjengelege og søkbare databasar og nettbasert presentasjon av det som finst. Det kom også fram eit sterkt ønske om nærare samarbeid mellom fagfolk og brukarar for nettopp å sikra høg kvalitet (og større grad av standardisering av terminologien på ulike fagområde). Dette fell saman med det translatørane treng: Skal ein omsetja tekst frå bestemte fagområde, skulle eg tru mange vil ha stor nytte av tilgang til ein termbase dei visste var kvalitetssikra. Draumen er basar av typen KB-N for alle fag.

### **Kva kan translatørane gje attende?**

I ei ideell verd ser eg for meg at det er laga terminologibasar for ulike fag, og at desse er tilgjengelege for ulike brukarar, anten som lisensierte fagbasar på line med andre språkressursar og til bruk i omsetjingsverktøy eller som basar ein kan gjera søk i (med brukarnamn og passord). Translatørane vil naturleg vera ei stor brukargruppe, men før vi får den ideelle verda, må mykje skje. Translatørar er ei faggruppe som sit med stor kunnskap om faget sitt og områda dei omset til/frå. Denne kunnskapen vil eg gjerne at dei deler med andre og gjerne i form av å bidra til språkbanken med terminologilister som dei sjølve eller andre kan tilordna og setja inn i vidare samanhengar. Ein språkbank vil ha ulike verktøy tilgjengeleg, og verktøy for ekserpering og systematisering av terminologi bør vera blant desse.

### **Språkbanken er infrastruktur for språkteknologisk utvikling**

Språkbanken kan bli ein sentral for systematisk utveksling av fagkunnskapar og terminologi i tillegg til den generelle funksjonen han vil ha som leverandør og formidlar av generelle språkressursar. Språkbanken skal ikkje koma i staden for andre organ eller organisasjonar, men heller bli eit koordinerande tillegg som gjer at fleire miljø kjem i kontakt med kvarandre og kan byggja på kvarandre sine kunnskapar og ressursar. Eg ser for meg at det er ein inngang til språkbanken, innhaldet kan liggja andre stader. Inngangen til språkbanken gjev informasjon om tilgjengelege språkressursar og korleis ein skal gå fram for å få tilgang til dei. Språkbanken vil stå for å klarera alt rundt opphavsrettar og bruksrettar, og det vil vera språkbanken som gjev brukarane tilgang til aktuelle språkressursar og sørgjer for at mottakaren får det han eller ho har bede om.

### ***Avsluttande merknad***

Etter at artikkelen vart skriven våren 2008 har regjeringa lagt fram St.meld. nr. 35 (2007-2008) *Mål og meining*. I meldinga står det at språkbanken skal etablerast frå 1. januar 2009<sup>11</sup>.

## Litteratur

- Berkenkotter, Carol & Huckin, Thomas N. (1995): Da klokka klang ... Om å lære seg fagsjangerer i skole og på universitet i Evensen, Lars Sigfred og Hoel, Torlaug Løkensgard: *Skriveteorier og skolepraksis*, LNU / Cappelen Akademisk Forlag 1997, s. 112–129.
- eNorge handlingsplan* (2000), versjon 1.0, Nærings- og handelsdepartementet.
- Jonassen, Knut (2005): Norsk terminologi. Kartlegging av nasjonale fagmiljøer, infrastruktur og holdninger. Prosjektrapport i Hoel, Jan (red.): *Hvem tar ansvaret for fagterminologien?* Rapport fra en strategikonferanse om terminologi og fagspråk i Norge, Oslo: Språkrådet 2005. s. 68–126.
- Krauwer, Steven /ELSNET (2003): *BLARK (The Basic Language Resource Kit) as the First Milestone for the Language Resources Roadmap*, <http://www.elsnet.org/dox/krauwer-specom2003.pdf>
- Kristiansen, Marita (2006): Fagspråkstrategien til Noregs handelshøgskole i *Språknytt* 2/2006.
- Noregs forskingsråd (2001): *KUNSTI-programmet*, <http://program.forskningsradet.no/kunsti/no/>
- Norsk språkråd (2001): *Handlingsplan for norsk språk og IKT*.
- Norsk språkråd (2002): "Samling og tilgjengelegging av norske språkteknologiressursar", prosjektplan for norsk språkbank.
- Stortingsmelding nr. 48 (2003-2004) *Kulturpolitikk fram mot 2014*, kap. 12.9: Ein norsk språkbank.
- Stortingsmelding nr. 35 (2007-2008) *Mål og meining*, kap. 7.5.6.2 Ein norsk språkbank skal etablerast
- Svensen, Torbjørn, m.fl. (1999) *Norsk språkbank, utredning om et nasjonalt korpus for språkteknologi*. NTNU 1999

---

<sup>11</sup> St.meld. nr. 35 (2007-2008) *Mål og meining*, s. 136