SAMLING OG TILGJENGELEG-GJERING AV NORSKE SPRÅK-TEKNOLOGIRESSURSAR

Prosjektgruppe oppnemnd av
Kultur- og kyrkjedepartementet
Rapport, oktober 2002

Forord

Prosjektgruppa oppnemnd av Kultur- og kyrkjedepartementet 19.3.2002 leverer med dette den endelege rapporten som svar på oppdraget ho fekk med å greie ut sentrale problemstillingar innanfor feltet samling og tilgjengeleggjering av norske språkteknologiske ressursar.

Det er konsensus i prosjektgruppa om innhaldet i rapporten. Alle tilrådingar og konklusjonar er uttrykk for gruppa si eiga oppfatning, men gruppa vil rette ein stor takk til den ressursgruppa som har representert ulike brukarmiljø. Ressursgruppa har kome med mange verdfulle og nyttige innspel undervegs i arbeidet.

Tilhøvet til samisk var ikkje med i mandatet for arbeidet, men prosjektgruppa er kjend med at det er sett i gang eit arbeid for å etablere ein samisk språkbank (Strategi for elektronisk innhold, NHD april 2002).

Torbjørn Svendsen
leiar (til juni 2002)

Torbjørn Nordgård
nestleiar (leiar frå juni 2002)

Leiv Hartly Andreassen

Jon Trygve Berg

Knut Kvale

Tron Espeli

Stig Johansson

Torbjørg Breivik
sekretær

OPPNEMNING, MANDAT OG ARBEIDSMÅTE

Kultur- og kyrkjedepartementet oppnemnde i brev av 19.3.02 denne prosjektgruppa:

Professor Torbjørn Svendsen, NTNU (leiar)
Professor Torbjørn Nordgård, NTNU (vara for Svendsen og nestleiar i prosjektgruppa)
Rådgjevar Tron Espeli, Noregs forskingsråd
Adm.dir. Leiv Hartly Andreassen, SAIL Port Northern Europe AS (SPNE)
Seniorforskar, professor Knut Kvale, Telenor ASA
Product Planner Signe Marie Flåt, Microsoft Norge AS
Professor Stig Johansson, Universitetet i Oslo

Alle medlemmene hadde personlege varamedlemmer som har teke aktivt del i arbeidet:
Rådgjevar Bernt-Erik Heid, Noregs forskingsråd (for Espeli)
Avdelingsleiar Anja Hilt / teknisk sjef Jon Trygve Berg, Nordisk språkteknologi AS (NST) (for Andreassen)
FoU-sjef Robert Engels, CognIT a.s. (for Kvale)
Professor Helge Dyvik, Universitetet i Bergen (for Johansson)

Direktør Bente Maegaard, Center for sprogteknologi, København, og professor Lars Ahrenberg, Universitetet i Linköping, har vore eksterne observatørar/ressurspersonar.

Signe Marie Flåt har seinare trekt seg frå arbeidet. Jon Trygve Berg har fungert som fast medlem i prosjektgruppa.

På oppdrag frå Kultur- og kyrkjedepartementet oppnemnde Språkrådet denne ressursgruppa for arbeidet:

Leiar Jan Olav Fretland, Norsk språkråd (leiar for gruppa)
Rådgjevar Grete Knudsen, GBM-Partners, Bergen
Generalsekretær Per Morten Hoff, IKT-Norge, Oslo
Dagleg leiar Bjørn Seljebotn, Nynodata a.s, Bø i Telemark
Generalsekretær Trond Andreassen, Norsk faglitterær forfatter- og oversetterforening, Oslo
Direktør Øyvind Haaland, Berlitz GlobalNET, Bergen
Dagleg leiar Ove Nyland, Noreg.no, Leikanger
Underdirektør Petter Korseth, Læringssenteret, Oslo
Koordinator for språkteknologi Kristin Bech, HIT-senteret / Universitetet i Bergen
Kredittdirektør Bjørn Norman Hansen, SND, Oslo
Professor Ruth Vatvedt Fjeld, Universitetet i Oslo

Sekretariat:
Rådgjevar Torbjørg Breivik har fungert som sekretær for arbeidet i begge gruppene.
Rådgjevar Marit Hovdenak deltok i sluttføringa av Delrapport 1.

MANDAT

”Samling og tilgjengeliggjøring av norske språkteknologiressurser

MÅL
Klargjøre forutsetningene for å samle og gjøre tilgjengelig norske språkteknologiressursar gjennom en egen utredning**.**

DELMÅL 1
Få fram et beslutningsgrunnlag for budsjettprosessen i 2003 innen 15. juni 2002.

DELMÅL 2
Ferdig utredning med forutsetninger for og forslag til hvordan en norsk språkbank kan etableres og utvikles. Frist for sluttrapporten: 30. september.

MANDAT FOR ARBEIDET

1 Framskaffe en oppdatert og realistisk analyse av behovet for å gjøre norske språkressurser tilgjengelig og hvilken innsats dette fordrer. Dersom tidligere behovsvurderinger blir opprettholdt, bør det spesifiseres hvilken type ressurser som trengs, og i hvilket omfang, til henholdsvis forskning og industriformål. Det bør defineres et minimumsbehov samt ønskelig omfang, og utarbeides kravspesifikasjoner for de ulike scenarioer.

2 Utrede juridiske problemstillinger knyttet til innsamling, leveranser og utvikling av språkmateriale.

3 Utrede datatekniske problemstillinger: valg av teknologi for drift, tilrettelegging og leveranse av materiale. Her må man ta hensyn til internasjonale standarder/valg for tilsvarende baser. Verktøy for sortering og leveranse må også være fleksibelt nok til å ivareta kravene fra ulike mottakere (forskning og industri).

4 Skissere en organisasjonsmodell som ivaretar prinsipp som uavhengighet (selvstendig juridisk person), bemanning, krav til kompetanse i organisasjonen.

5 Skissere en driftsmodell som sikrer at brukergruppene (forskning og industri) kan nyttiggjøre seg ressursene gjennom de tekniske løsningene som velges, de juridiske avtalene som utarbeides, og den finansierings-/betalingsmodellen som foreslås.

6 Utarbeide en finansieringsplan. Denne planen må skissere hva det koster å etablere selve banken, hva det koster å skaffe og legge inn så mye ressurser at relevante brukergrupper kan nyttiggjøre seg innholdet (kritisk suksessfaktor), en opptrappingsplan for komplett samling, og hva det koster å drifte og utvikle selve språkbanken. Økonomien baseres på at dette blir en kombinasjon av privat og offentlig finansiering. Finansieringskilder identifiseres. Tiltakets sektorovergripende karakter skal gjenspeiles i de finansieringsmodeller som foreslås.”

ARBEIDSMÅTE

Prosjektgruppa har hatt seks møte. Ressursgruppa har hatt tre. Mellom møta har medlemmene i begge gruppene kome med innspel. Prosjektgruppa har gjeve dei ulike punkta i mandatet ulik vekt. Den juridiske utgreiinga blei sett bort til advokatfirmaet Simonsen Føyen DA og følger som eige vedlegg. Punkt 3 i mandatet om datatekniske problemstillingar reiser ikkje særskilde prinsipielle problemstillingar fordi det på fleire område alt finst internasjonale standardar som må følgjast, og der det ikkje er tilfelle, bør ein leggje seg på det som synest å ha etablert seg som praksis. Gruppa har prioritert å bruke tida på dei andre punkta. Prosjektgruppa har gjennomført to studiereiser: til ELRA/ELDA*, Paris, og til NST** og SPNE***, Voss. Rapportane frå desse reisene følgjer som vedlegg til rapporten. Alle dokument og innspel har vore tilgjengelege for medlemmene i begge gruppene på ein eigen nettstad.

Delmål 1 og oppgåve 2 i mandatet blei leverte 18. juni 2002 ved at Delrapport 1 blei overrakt av prosjektgruppa saman med ei juridisk utgreiing frå Simonsen Føyen.

Konklusjonane og det meste av innhaldet i Delrapport 1 er tekne inn i denne rapporten, som elles gjev meir utførleg omtale av dei fleste punkta i Delrapport 1. Kartlegging av kva som finst, omtalen av kva ressursar som trengst, og utarbeiding av ein finansieringsplan og forslag til gjennomføring er det gruppa har konsentrert innsatsen om i sluttrapporten.

*European Language Resource Association / European Language Distribution Agency
**Nordisk språkteknologi AS
***S.AI.L. Port Northern Europe AS

INNHALD

KAPITTEL 1 KONKLUSJONAR

Mandatet gjev prosjektgruppa i oppdrag å greie ut "Samling og tilgjengeleggjering av norske språkteknologiske ressursar". Gruppa har valt å nytte termen *språkbank* om ei slik samling. Ein språkbank skal forvalte ein kapital, ein nasjonal språkressurs. Like lite som ein bank har all kapitalen sin i éin, sentral kvelv, treng språkbanken å vere lokalisert til éin einskild stad. Termen ”språkbank” er òg vel innarbeidd i alle aktuelle fagmiljø.

Hovudkonklusjonen er at det er særs viktig å få etablert norske språkteknologiske ressursar så snart som råd er. Denne oppfatninga deler prosjektgruppa og ressursgruppa med eit samla forskings- og industrimiljø innanfor språkteknologien i Noreg. Med ei slik samling kan ein

- medverke til å oppfylle målsetjinga om at norsk – i tale og skrift – skal vere det dominerande bruksspråket i det norske samfunnet
- sjå til at norsk språkteknologi medverkar til å fremje samfunnsdeltaking og kulturell identitet gjennom å ta i bruk den samla norske språkkulturen
- styrkje det norske språket (=bokmål, nynorsk og norske dialektar) og motverke domenetap, dvs. at engelsk gradvis tek over som bruksspråk på fleire område
- stimulere norsk IKT-industri til å satse på språkteknologiske løysingar for norsk og andre språk
- utnytte det verdiskapingspotensialet som ligg i å kople høg IKT-kompetanse med høg kompetanse om språk
- gjere det attraktivt for utanlandske leverandørar å lage norskspråklege produkt

Fleire land i Europa er i ferd med å etablere ein språkteknologisk infrastruktur lik den som blir foreslått her, primært for å styrkje nasjonalspråka si stilling mot ein aukande påverknad frå engelsk. Dessa landa har frå før eit langt betre utgangspunkt enn Noreg fordi dei alt i 1990-åra fekk på plass ein del relevante ressursar. Alle desse initiativa har ei vesentleg grunnfinansiering frå det offentlege, og i Nederland medrekna den flamske delen av Belgia er innsamlinga 100 % offentleg finansiert.

Ressurssamlinga bør organiserast som ei stifting med offentleg eigarskap, eit breitt samansett styre og minimal administrativ bemanning. Innsamling, drift, vedlikehald og distribusjon bør setjast bort til eksisterande aktørar med relevant kompetanse og erfaring.

Språkbanken bør vere ått av ei stifting som er under offentleg kontroll, for at ein kan sikre klare eigartilhøve og bruksrettar. Det er uheldig med ei samanblanding av privat og offentleg eige, både av juridiske grunnar og av omsyn til konkurransen i marknaden. Realistisk sett må nesten all finansieringa vere offentleg. Slik er det i land som har sett i gang liknande prosjekt. Den relevante norske industrien har ikkje styrke til å kunne vere med på finansieringa i særleg grad. Dette ser vi også gjeld internasjonalt.

Samlinga må innehalde materiale som er tenleg for språkteknologisk industri og forsking, representativt materiale for ulike bruksområde (tekst, tale, dialektar, begge målformer osv.), godt dokumentert materiale som er koda i høve til internasjonale standardar. Materialet må vere kvalitetssikra (validert), og alle bruksrettar må vere avklarte. Ei nasjonal ressurssamling trengst for at norsk språk skal bli godt representert i IKT-løysingar der naturleg språk er ein del. Utan gode norskspråklege produkt og tenester vil engelsk bli meir og meir dominerande i næringsliv, utdanning og forvaltning.

KAPITTEL 2 KVIFOR EI NORSK RESSURSSAMLING?

Språkteknologi forenklar og forbetrar språkleg kommunikasjon. Mange produkt og tenester finst allereie for engelsk, men i langt mindre grad for norsk. Skal ein ta vare på norsk språk, kultur og identitet, må ein skape eit grunnlag for norske språkteknologiske produkt. Dette grunnlaget må vere i form av ei samling språkdata av tilstrekkeleg storleik og kvalitet.

Språk er ein føresetnad for samfunnet vårt. Kommunikasjon mellom menneske er bygd på språk, i munnleg og skriftleg form. I dag er det i aukande grad mogleg å bruke naturleg språk i kommunikasjonen mellom menneske og maskin.

2.1 Kvifor språkteknologi?

Informasjons- og kommunikasjonsteknologien, IKT, blir stadig viktigare i samfunnet. Bruk av Internett og datamaskiner er ikkje i bruk berre i heim og på arbeid. IKT inngår meir og meir i så godt som all teknologi, frå hushaldningsapparat, forbrukarelektronikk og bilar til profesjonell og tung teknologi. I stadig fleire samanhengar er det ikkje mangelen på informasjon eller på innebygde finessar som gjev brukarane problem, problemet er å velje ut det som er interessant.

Informasjonsutveksling inneber ofte at ein må nytte elektroniske medium til lagring og overføring. Gode verktøy for dokumentgenerering og -redigering er ein føresetnad for effektiv informasjonsutveksling.

Språkteknologi dreier seg om å forenkle og forbetre kommunikasjonen mellom menneske, og om å gjere samvirket mellom menneske og maskin enklare. Denne teknologien er med på å gjere det lettare å nytte moderne informasjonsteknologi fordi brukarane kan kommunisere på den måten dei er mest fortrulege med, nemleg med tale- og skriftspråket sitt. Fleire kan få tilgang på informasjon, tenester og produkt fordi terskelen for å nytte informasjonsteknologien blir lågare. Døme på språkteknologi er automatisk talegjenkjenning (datamaskinen omset tale til tekst) og maskinprodusert tale, maskinomsetjing og hjelpemiddel til dokumentproduksjon og informasjonsgjenfinning.

Språkteknologi kan effektivisere mange arbeidsprosessar. Eit par enkle reknedøme kan illustrere dette, og vi hentar det første frå sjukehusverda. Meir enn 3000 årsverk er relaterte til skriving av journalar etter diktat, og lønnskostnadene er på meir enn kr 300 000 pr årsverk. Det betyr ein årleg kostnad på om lag ein milliard kroner. Erfaringar frå Philips viser at sekretærar kan produsere rapportar opp til 40 % raskare ved å bruke dikteringsverktøy. Dersom ein kan få ein effektiviseringsgevinst på 10 % knytt til diktering ved sjukehusa, noko som er eit svært nøkternt overslag, dekkjer ein inn utgiftene til språkbanken sitt minsteomfang på berre eit driftsår ved sjukehusa. Det er god grunn til å rekne med at ein vil spare mykje meir når ein tek med at diktering etter kvart kan takast i bruk på mange andre område av offentleg sektor, vel å merke dersom det finst språkdata til å trene systema opp med. I tillegg kjem innsparingar i privat sektor.

Eit anna døme er gjenfinning av informasjon som eksisterer i elektronisk form. Eit gjennomgåande problem er at informasjon blir utilgjengeleg når informasjonsmengda aukar. Språkteknologi kan bøte på dette ved å automatisere informasjonsindeksering og informasjonsgjenfinning (”digitale bibliotekarar”).

Eit tredje døme er automatisk omsetjing mellom engelsk og norsk, eller datastøtta omsetjing mellom desse to språka. Mengda av omsette tekstar både i offentleg og privat sektor er svært stor, og mykje ressursar går med til manuell omsetjing. Dette er arbeid som må utførast av høgt kvalifisert personale for at til dømes teknisk og juridisk informasjon skal bli omsett korrekt. Det er vanleg å rekne med 20–40 % mindre tidsbruk ved datastøtta maskinomsetjing. Om ein går ut frå at det blir brukt eit par tusen årsverk årleg til omsetjing mellom norsk og engelsk, vil relevante verktøy føre til store innsparingar i offentleg så vel som i privat sektor. Investeringane i språkbanken vil raskt bli tente inn.

Mange språkteknologiske produkt og tenester er alt tilgjengelege på den internasjonale marknaden. Dei fleste finst diverre ikkje på norsk, korkje på bokmål eller på nynorsk. Grunnen er mellom anna manglande språkressursar. Engelsk er det dominerande språket. For å kunne utnytte fordelane med språkteknologien må brukarane kunne bruke morsmålet sitt. Først då blir teknologien tilgjengeleg for alle.

Noreg har lågt folketal samanlikna med til dømes Tyskland, Frankrike og Storbritannia. Det gjer at marknaden for norskspråklege produkt er liten, og svært mykje mindre enn for dei store europeiske språka. Det set òg grenser for kor mykje av kostnaden med å etablere den ønskte databasen med norske språkdata ein kan vente at industrien vil finansiere, fordi kostnadene med å utvikle språkteknologiske produkt er omtrent den same for alle språk. Innhaldet i ei slik ressurssamling som språkbasen vil bli, vil i tillegg vere til stor nytte for den generelle språkforskinga. Billeg tilgang på grunnleggjande språkressursar er avgjerande for norske aktørar og for at internasjonale aktørar skal sjå Noreg som ein interessant marknad.

Språkteknologiindustriens mål er, som for all anna kommersiell verksemd, å utvikle lønsame produkt, dvs. produkt som marknaden ønskjer og treng. Vi opplever i dag språkteknologiske produkt og applikasjonar som med talekommandoar og talegjenkjenning løyser hushaldningsoppgåver, effektiviserer, sikrar og betrar framkome og flyt i trafikken (navigasjonssystem), set om mellom ulike språk (maskinomsetjing), og på fleire område gjer kvardagen lettare for folk. Ikkje minst gjeld dette for dei med ulike funksjonshemmingar.

Situasjonen er likevel slik at desse produkta i det store og heile berre er tilgjengelege på engelsk eller andre større språk. Språkteknologisk industri kan framstille desse produkta slik at dei kan fungere for alle her i landet, utan at ein er avhengig av kva dialekt brukarane har eller kva miljø produkta blir brukte i. Føresetnaden er at norske språkteknologiske ressursar blir gjort tilgjengelege. Norsk vil på denne måten kunne nyttast på lik line med engelsk innanfor eit teknologisk område som blir svært viktig for brukarane i framtida.

Det er eit overordna mål å sikre norsk som bruksspråk i alle samanhengar i Noreg. Det inneber at vi må leggje best mogleg til rette for at nordmenn skal kunne halde fram med å kommunisere med kvarandre på norsk. Morsmålet er det språket vi uttrykkjer oss best på, og som vi forstår best. Engelsk er andrespråket vårt. Når det finst verktøy for engelsk språk, men ikkje for norsk, kan det fort forsterke tendensen til å velje engelsk framfor norsk, til dømes som internt bedriftsspråk.

Språk er kultur og identitet. Det opplever vi i alle sosiale samanhengar – i heim, skule, arbeid og fritid. Datamaskinar er i ferd med å bli ein ny ”partnar” som vi kommuniserer med i aukande grad. Det vil vere eit alvorleg kulturpolitisk nederlag om norsk språk blir fortrengt av engelsk fordi ein ikkje ser seg råd til å leggje til rette for norsk språkteknologi.

2.2 Språkdata

Utvikling av språkteknologi krev teknologikunnskap, språkkunnskap og digitale språkressursar. Med få unntak nyttar all moderne språkteknologi seg av statistiske modellar av ymse slag. Til dømes vil eit dikteringssystem som automatisk konverterer tale til tekst, nytte statistisk modellering av uttalen av språklydar, og av samanhengen mellom ord. Desse modellane må trenast opp ved at ein bruker døme på tale og tekst frå store databasar. Trening av statistiske modellar krev eit mykje større tekstgrunnlag enn det som trengst til dømes for å skrive ein tradisjonell grammatikk eller ordbok. Treningsfasen er den mest sårbare delen når systemet skal lagast. Dersom treningsgrunnlaget, dvs. treningsdata, er for små eller av slett kvalitet, blir produktet dårleg. Dette generelle poenget om tilhøvet mellom data og produktkvalitet gjeld ikkje berre for dikteringssystem, men òg for maskinomsetjing, talesyntese, korrekturprogram osv.

Det største hinderet for å realisere målet om at alle innbyggjarane skal få lik tilgang til og høve til å ta i bruk den nye teknologien, er mangelen på norske språkdata. Språkdata på norsk trengst for at språkindustrien skal kunne utvikle *norskspråklege* produkt. Kravet til mengd og kvalitet på språkdata er stort sett uavhengig av kva språk det er, slik at kostnaden med å etablere språkressursar for norsk vil vere like stor som for til dømes engelsk. På grunn av den norske språksituasjonen med to målformer som begge har stor valfridom, og stor aksept for dialektbruk, vil kostnaden vere høgare enn for andre europeiske språk det er naturleg å samanlikne med.

2.3 Internasjonalt

Mange andre land i Europa har sett faren for at deira eige språk skal bli skadelidande om dei ikkje sjølve syter for å etablere dei språkressursane som må til for at innbyggjarane skal få dei nye tenestene på sitt eige språk. EU har prioritert språkleg mangfald, og støtta i 1990-åra fleire slike datainnsamlingar gjennom rammeprogramma, anten som reine datainnsamlingsprosjekt eller i tilknyting til forskingsprosjekt.

ELRA/ELDA (European Language Resource Assosiation / European Language Distribution Agency), Paris, er ein organisasjon som formidlar språkressursar, primært innanfor EU-landa, men organisasjonen samarbeider med andre liknande organisasjonar (til dømes Linguistic Data Consortium i USA). ELRA er ein medlemsorganisasjon. Medlemsorganisasjonen kan ikkje selje noko, men har stifta distribusjonsselskapet ELDA som marknadsfører ressursane og tenestene via ein katalog på nettet (http://www.elda.fr/cata/tabtext.html). Organisasjonen har ein avtale med EU-kommisjonen som tilseier at alle som får pengar frå EU til prosjekt som inkluderer innsamling av språkmateriale, må stille dette til rådvelde for ELRA/ELDA. Opphavs- og eigedomsretten til språkressursane ELDA formidlar, blir verande hos dei som har samla inn ressursane. ELDA syter for uavhengig validering av ressursane og fører royalties tilbake til eigar ved sal. Organisasjonen blir driven som ein non profitorganisasjon.

Katalogen til ELDA inneheld tekstkorpus, talekorpus og leksikalske data for ulike språk, ikkje alle typar ressursar er tilgjengelege for alle språka. Mange av ressurssamlingane er fleirspråklege, og katalogen inneheld også ikkje-europeiske språk, som til dømes japansk og kinesisk. Norsk er representert med ein mindre del av eit større fleirspråkleg talekorpus (SpeechDat) ved at Telenor tok del i dette EU-finansierte prosjektet. Av dei større korpusa kan nemnast British National Corpus (BNC) med 100 millionar ord og European Corpus Initiative (ECI) med 98 millionar ord (fleirspråkleg).

I dei fleste land der ein har sett i gang innsamling av språkressursar, er arbeidet eit resultat av samarbeid mellom fleire sektorar: nasjonale styresmakter, forsking og (språk)industriverksemder. Av noverande internasjonale innsamlingsinitiativ vil vi trekkje fram Italia, Frankrike og Nederland/Belgia. Italia og Frankrike legg opp til om lag 50 % offentleg finansiering. Dette er i begge landa ei utviding av eksisterande språkdata i form av ny innsamling av tekst og tale til ei samla kostnadsramme på om lag 8 millionar euro. Nederland og Flandern har til saman i overkant av 20 millionar nederlandsktalande. Her blir det samla inn taledata med 100 % offentleg finansiering til ein totalkostnad av knapt 5 millionar euro. Denne datasamlinga samsvarer godt med det som er identifisert som behovet for norske taledata. Ein viktig motivasjon for at denne innsamlinga blei fullfinansiert av det offentlege, var at ein ville sikre klare eigedomstilhøve.

Ingen av dei nordiske landa har noka samla, nasjonal språkressurssamling til bruk i forsking og industri. Mange institusjonar har kvar for seg og til eigne formål samla inn, vidareforedla og lagra nokre elektroniske språkressursar. For det meste gjeld dette universitets- og forskingsmiljø som arbeider med språkteknologi og datalingvistikk. Samlingane ved Göteborgs universitet, som består av eit svensk talemålskorpus og ein ”Svensk språkbank” (tekstdata, leksikalske ressursar) er kanskje det som ligg nærast opp til ei nasjonal ressurssamling for språkforsking i Norden.

2.4 Språkteknologi i det politiske biletet

I eNoreg 2005 har regjeringa tre overordna visjonar for IT-politikken: verdiskaping i næringslivet, effektivitet og kvalitet i offentleg sektor og deltaking og identitet (*eNorge 2005*, NHD mai 2002). Statsminister Kjell Magne Bondevik understreka desse punkta og utdjupa dei med å peike på kor viktig det er å ta vare på norsk språk, kultur og identitet i talen han heldt på eit seminar om planen (Oslo, 12.6.2002).

Stortingsmeldingar dei seinare åra har understreka at noko må gjerast for å demme opp for påverknaden frå andre språk, og særleg engelsk innanfor IKT-området. *St.meld. nr. 13 (1997–1998) Målbruk i offentleg teneste* la grunnlaget for å opprette IKT og språk som eiga avdeling i Norsk språkråd (s. 29). Same meldinga meiner det er nødvendig å lage ein eigen handlingsplan for norsk språk og IKT (s. 29). Neste melding om målbruk i offentleg teneste, *St.meld. nr. 9 (2001–2002)*, følgjer opp med ”Med dei utfordringane vi ser i kjølvatnet av internasjonaliseringa og den teknologiske utviklinga, må vi i tillegg stilla opp som det overordna språkpolitiske målet å verna og styrkja norsk språk, slik at både bokmål og nynorsk kan bestå som fullverdige bruksspråk i alle delar av samfunnslivet, inn i det nye informasjonssamfunnet.” (pkt. 1.2).

I første utgåve av eNoreg-planen (juni 2000) står Kulturdepartementet oppført som ansvarleg for at ein slik plan blir laga. Eit overordna mål i eNoreg, 1.0, er at alle skal ha lik tilgang til og like sjansar til å bruke den nye teknologien. I *Strategi for elektronisk innhald 2002–2004* (NHD, april 2002) er det eit mål å ha ”god tilgang på elektronisk kvalitetsinnhald som er produsert i Noreg eller lagt til rette for norske forhold”. Vidare seier planen: ”God tilgang til teknologiske språkdata gjer utviklinga av nye produkt langt meir kostnadseffektiv” (kap.1). I kapittel 9 i same planen blir det foreslått fleire pilotprosjekt, mellom anna eitt om språkteknologi, som denne rapporten er ein del av.

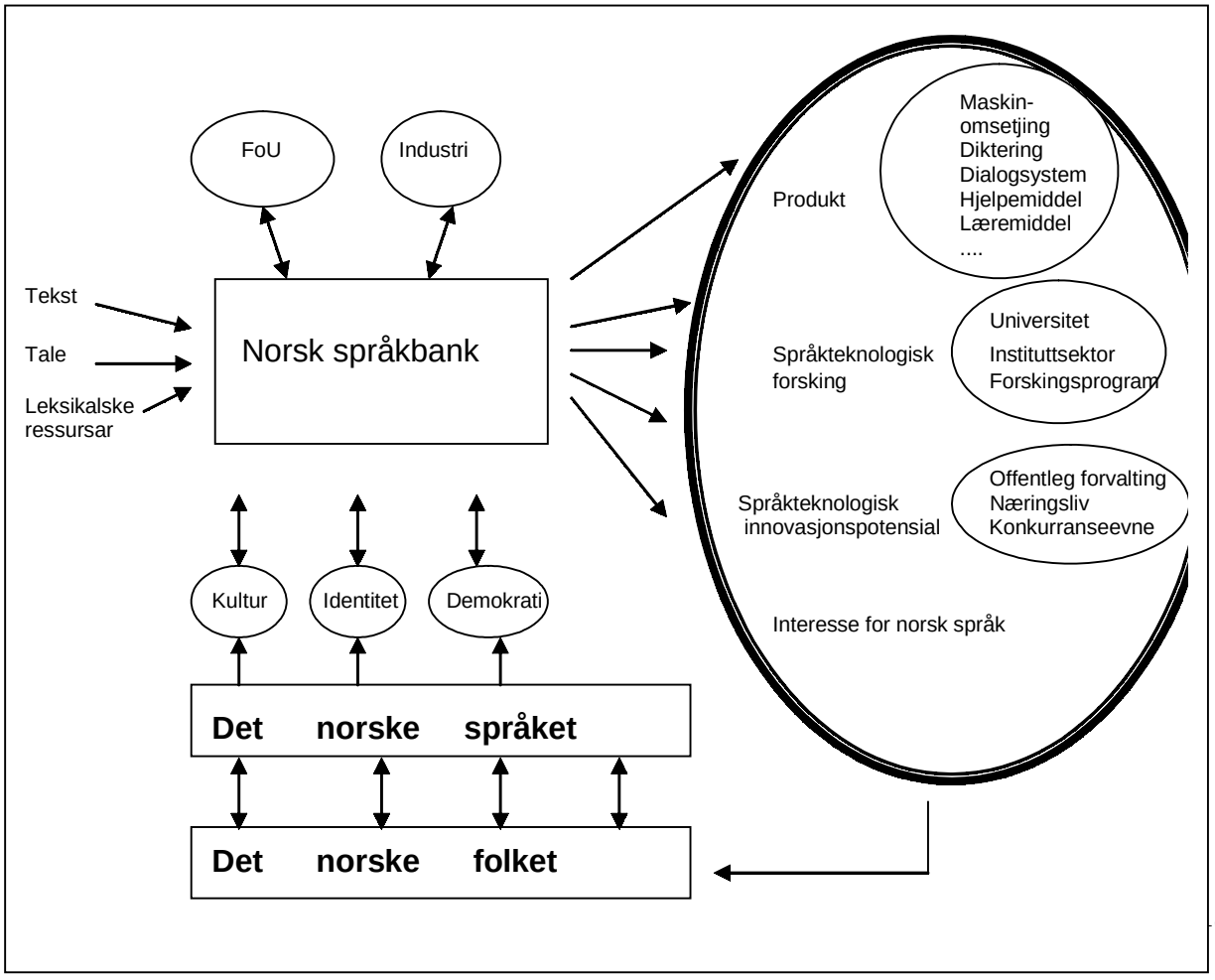
EU vurderer situasjonen tilsvarande og prioriterer no minoritetsspråk i det sjette rammeprogrammet. Mange av landa i EU har sett i gang innsamling av språkdata nettopp ut frå det same motivet: å sikre sitt eige språk som bruksspråk på alle område i samfunnslivet, ta vare

på sin eigen kultur og sin eigen identitet og samtidig demme opp for engelsk påverknad, særleg på IKT-området. I forslaget til handlingsprogram for det svenske språket, ***Mål i mun*** (SOU 2002:27), seier dei: ”Utvecklingen inom språkteknologin innebär både möjligheter och risker för svenska språket” (kap. 17.1.7). Den svenske utgreiinga viser elles til den norske *Handlingsplan for norsk språk og IKT*, og foreslår at Sverige må byggje opp ein nasjonal språkbank på same grunnlag som det som er blitt tilrådd i Noreg. I 1999 oppretta det danske Forskningsministeriet ei arbeidsgruppe for ”IT på dansk”. Arbeidsgruppa skulle ”rådgive Ministeriet for Videnskab, Teknologi og Udvikling omkring udvikling av IT på dansk, herunder udarbejdelsen av en dansk sprogteknologisk ordbase”, tema ”Sprog- og Tale-teknologi”. Ei tilråding frå gruppa blei levert i juli 2001. Denne har resultert i at STO-prosjektet (ein dansk ordbank) har kome i gang, men det er ikkje etablert nokon dansk språkbank tilsvarande den norske som prosjektgruppa her tilrår.

2.5 Ein norsk språkbank

Innsamling av norske språkdata er ein føresetnad for språkteknologiske produkt og tenester på norsk. Ei slik datasamling må vere tilgjengeleg for alle aktørar på den norske marknaden. Dette er eit nasjonalt kultur- og næringspolitisk ansvar. Sett i lys av den avgrensa marknaden må hovuddelen av kostnadene dekkjast av det offentlege. Etablering av norske språkressursar vil, saman med den satsinga Noregs forskingsråd gjer i språkteknologi, gje grobotn for ein norsk språkteknologisk industri og for tilgang på norskspråklege produkt og tenester. Språkressursar vil, saman med FoU-verksemd og språkteknologisk industri, kunne gje det norske samfunnet dei norskspråklege produkta og tenestene det er behov for.

Teikninga *(Figur 1)* nedanfor viser korleis prosjektgruppa ser for seg at språkbanken kan fungere i høve til relevante aktørar.



KAPITTEL 3 ORGANISERING

Ressurssamlinga bør organiserast som ei stifting med offentlege stiftarar, eit breitt samansett styre og minimal administrativ bemanning. Innsamling, drift, vedlikehald og distribusjon bør setjast bort til eksisterande aktørar med relevant kompetanse og erfaring.

3.1 Innleiing

Ein språkbank vil ha to typar oppgåver. På den eine sida skal han syte for at ressursane er av rett type og mengd. På den andre sida har han eit ansvar for å forvalte investeringa og syte for at ressurssamlinga kjem til nytte for aktuelle industri- og forskingsformål. Den første typen oppgåve er avgrensa til den tida det tek å etablere samlinga, medan den andre er permanent. I oppbyggingsfasen skal språkbanken ha ansvar for

- å frikjøpe eksisterande materiale og leggje det til rette for vidareformidling
- å identifisere materiale som kan vidareformidlast (utan vilkår om frikjøp)
- å organisere nyinnsamling av materiale

Han må òg sjå til at relevant tilpassing av eksisterande materiale blir gjord etter fagleg forsvarlege spesifikasjonar.

Ei viktig oppgåve er å syte for vedlikehald og vidareutvikling av ressursane. Materiale som ikkje blir halde ved like og utvikla vidare, tapar verdi over tid. Blant dei permanente oppgåvene finn ein

- distribusjonsarbeid
- ansvar for forbetringar
- ansvar for å foreslå nyinnsamlingar, m.a. å foreslå korleis nyinnsamlingar kan finansierast etter at banken har nådd eit basisnivå
- ansvar for kvalitetskontroll i samband med vidareforedling og nyinnsamling

Det er viktig at ein ikkje bruker større innsats enn nødvendig til den administrative forvaltninga av språkressursane.

3.2 Eige eller formidle ressursane i samlinga?

Ei rein formidlingsrolle vil innebere at språkbanken har ansvar for å gjere språkressursar tilgjengelege for forskings- og industriformål, og slik at eigedomsretten ligg att hos den som har samla ressursane opphavleg. Dei som stiller materialet sitt til rådvelde, er sjølve ansvarlege for å ha avklart dei juridiske rettane til materialet. I fall misbruk blir oppdaga, er ikkje språkbanken den ansvarlege, men den som har stilt materialet til disposisjon.

Ein konsekvens av ei rein formidlarrolle er at språkbanken ikkje kan vidareforedle, velje ut eller leggje til rette ressursane på nye måtar. Ein kan tenkje seg kundar som treng spesifikke ressursar, og som ønskjer at andre vel ut og set saman desse ressursane for dei framfor at dei sjølve skal gjere det. ELRA/ELDA si erfaring er at formidlarrolla fører til at det blir mindre gjenbruk enn det kunne vore. Det taler til fordel for eigarskap eller ein vid bruksrett.

Satsar ein på ein språkbank som eig alle språkressursane sjølv, krev det at ein må gje kompensasjon til eigarane for eigedoms- og opphavsretten knytt til eksisterande materiale som kan inngå i samlinga. Språkbanken har då alle rettar til å vidareforedle, tilpasse og setje

saman materialet på nye måtar. Om ein skal løyse ut eksisterande materiale eller samle inn nytt, må vurderast i kvart einskilt høve ut frå kvalitet, gjenbruksverdi og pris i høve til kostnadene med nyinnsamling. Det kan vise seg å vere like dyrt og ressurskrevjande å leggje til rette eksisterande materiale for gjenbruk som det er å samle inn materiale frå grunnen av.

Ei juridisk utgreiing som prosjektgruppa har fått utført (vedlagd), avdekkjer at det kan vere vanskeleg å leggje delar av dei eksisterande språkressursane som ulike institusjonar har i dag, inn i ei nasjonal ressurssamling grunna manglande rettar til å bruke materialet i andre samanhengar enn den dei blei samla inn for. Særleg er det vanskar med materiale som er stilt til rådvelde for særskilde formål, noko som kan setje ein stoppar for gjenbruk. I slike høve kan ein gjere to ting: anten reforhandle eksisterande avtalar med kvar einskild som har gjeve frå seg materiale, eller samle inn nytt materiale.

Ein kombinasjon av det å eige og det å formidle vil vere fleksibel. Innsamla ressursar som kan gjerast tilgjengelege, vil ein då kunne formidle utan at det fører med seg ekstra kostnader (til dømes materiale som er samla inn for offentlege midlar ved universiteta), og ein har handlefridom til å setje i gang innsamling av nye ressursar. Ei ulempe vil vere at manglande eigarskap/bruksrett vil overlate til eigaren om ressurssamlinga skal vidareutviklast eller vidareforedlast. Det vil òg avgrense den strategiske styringa språkbanken kan utøve. Ideelt sett vil det vere ønskjeleg å ha eigarskap eller fulle bruksrettar til så mykje av språkressursane som råd er.

3.3 Innsamling

Språkbankorganisasjonen skal vere slank, effektiv og fleksibel. Det er ikkje ønskjeleg å byggje opp ein stor organisasjon som skal stå for innsamlinga av språkressursar, for så å avvikle hovuddelen av han etter at den store innsamlinga er ferdig. Organisasjonen skal heller vere oppdragsgjevar for innsamlingsarbeidet og få på plass ei fleksibel prosjektorganisering av desse oppgåvene på kostnadseffektivt vis. Eksisterande språkressursar må vurderast med omsyn til om dei samsvarer med behova til språkbanken, om dei har tilfredsstillande kvalitet, og om overtaking av eigedomsrettar og/eller bruksrettar vil vere kostnadssvarande samanlikna med nyinnsamling. Nyinnsamling av data bør gå føre seg etter anbod innanfor ein utlyst portefølje av ressursar som er prioriterte av styret for språkbanken.

Alle data som skal inngå i språkbanken, både eksisterande og nyinnsamla data, må validerast (kvalitetssikrast) av ein uavhengig institusjon i høve til dei som har gjort innsamlinga. Det må gjerast for å sikre at kvaliteten på språkressursane tilfredsstiller dei krava som er sette, og at innhaldet samsvarer med dokumentasjonen.

3.4 Distribusjon

Der det er mogleg, bør ein nytte eksisterande distribusjonskanalar slik at ein unngår å byggje opp ei unødvendig administrativ eining. Samtidig må ein sikre at dei som betaler for ressursoppbygginga, har reell eigarskap til ressursane.

Ressursar som skal vidareformidlast, kan vere lagra der dei ligg i dag, dersom det er mest praktisk. Andre ressursar som er frikjøpte eller nyinnsamla, kan vere lagra ulike stader. Det som er avgjerande, er at ressursane kan leverast raskt, og at det er tilgang på tilstrekkeleg kompetanse til å ta seg av det som blir kravd av administrasjon, drift og oppgåver knytte til vidareforedling, vedlikehald, tilrettelegging, kopiering og utsending.

I europeisk samanheng er ELRA/ELDA det viktigaste organet som formidlar språkressursar. Dei norske språkressursane må stillast til disposisjon for internasjonale brukarar gjennom ELRA/ELDA. Den norske språkbankorganisasjonen må samarbeide tett med ELRA/ELDA når det gjeld utvikling, bruk av standardar og krav til kvalitet.

3.5 Organisasjonsform

Prosjektgruppa ser to organisasjonsformer som peiker seg ut til å forvalte den norske ressurssamlinga: aksjeselskap og stifting. I diskusjonen om organisasjonsform kjem tilhøvet til ressursane inn: Skal organisasjonen berre formidle ressursar, vil han berre fungere som eit bindeledd mellom eigar og brukar. Det vil setje små krav til organisasjonsforma. Om organisasjonen skal stå som eigar eller ha vide bruksrettar til ressursane, stiller det seg annleis. Då må han vere ein juridisk person (eit sjølvstendig rettssubjekt).

Ved val av selskapsform er det viktig å ta omsyn til at organisasjonen må vere fleksibel, kunne handle raskt, og ha fullmakter som tilseier at ein kan setje ut oppdrag. Det er også eit vesentleg moment at han har eit breitt samansett styre der relevante interesser er representerte, det vil seie industri og forsking. Eigarinteressene må òg vere høveleg representerte i styret. Styret legg strategi og gjev prioriteringane. Styret må ha solid kompetanse når det gjeld behova i språkteknologisk industri og språkteknologisk forskings- og utviklingsarbeid, nasjonalt og internasjonalt.

Aksjeselskap og stifting er jamstelte juridisk når det gjeld handlefridom, rom for å auke kapitalen undervegs, offentleg innsyn i rekneskapane og langt på veg kva som skjer om organisasjonen blir nedlagd. Skilnaden ligg i at eit aksjeselskap har eigarar, medan ei stifting ikkje har eigarar. Stiftaren eller stiftarane stiller ein stiftingskapital til rådvelde for stiftinga. I situasjonar der ein står overfor nedlegging eller konkurs, skil dei seg: ressursane går attende til stiftarane om ein har ei stifting, medan kven som helst kan kjøpe eit konkursbu frå eit aksjeselskap.

I eit aksjeselskap er det eigarane som peiker ut eller vel styret. Dersom den eller dei som opprettar ei stifting, ønskjer styring med korleis stiftingskapitalen blir forvalta, må det sikrast gjennom vedtektene for stiftinga.

Prosjektgruppa meiner at aksjeselskap er problematisk fordi selskapet kan gå konkurs og samlinga dermed gå tapt som ressurs for fellesskapet.

Eit særlovsselskap kan vere ei løysing, men det tek tid å få lovverket på plass. I så fall må ein arbeide for det parallelt med etableringa og innsamlinga. Tidsaspektet er viktig: Arbeidet må setjast i gang så snart råd er.

Prosjektgruppa ser ei stifting med departement som stiftarar som det mest nærliggjande alternativet. Denne organisasjonsforma vil framheve at språkbanken har karakter av ei samling med nasjonale fellesressursar. Det vil òg gje meir stabile føresetnader for organisasjonen enn aksjeselskapsforma, og eliminere risikoen for at desse nasjonale fellesressursane kan gå tapt ved ein konkurs. Stiftinga må ha ein ”formuesverdi [som] blir stilt til selvstendig rådighet” dvs. at ho ved etableringa må ha ein grunnkapital. Det verkar naturleg at dei gjeldande departementa går inn med dei etableringsmidlane som skal til, som ei form for ”medgift” men at det blir opna for at andre også kan kome inn med midlar, jamfør kapitlet om finansiering.

Stiftinga sine vedtekter må gje klart uttrykk for formålet med samlinga, og definere ein styrings- og rådsstruktur som er eigna for dette. Styret må vere sett saman av dei viktigaste aktørane i høve til finansiering (departement) og representantar for dei viktigaste brukar-interessene (språkteknologiindustri, forsking, offentlege og private brukarmiljø).

Stiftinga må ha ei etter måten fri stilling i høve til departementa, som må skjøtte sine interesser gjennom styrerepresentasjon. Av praktiske grunnar bør stiftinga samlokaliserast med ein eksisterande institusjon, til dømes Norsk språkråd, for å minimalisere administrative kostnader (lokale, teknisk infrastruktur, generelle administrative støtteoppgåver) – særleg i tilknyting til etableringa.

Det må ikkje vere tvil om kven som har eigedoms- og bruksrett til ressursane som blir bygde opp. Det må lagast eigne avtalar som regulerer høvet mellom eigedoms- og bruksrett i fall språkbanken ikkje er eigar. Når organisasjonen formidlar eksisterande ressursar, vil ikkje det røre ved eigedomsretten.

3.6 Drift og vedlikehald

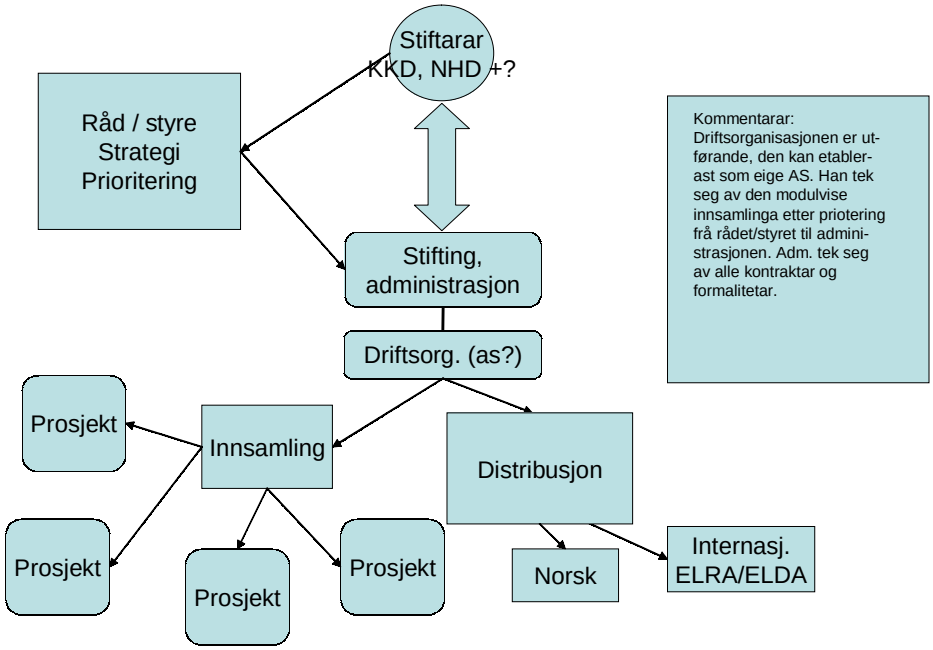
Ein kan tenkje seg ulike løysingar når det gjeld drifta av språkbanken. Prosjektgruppa ser for seg at stiftinga Norsk språkbank (organisasjonen som eig og forvaltar ressurssamlinga) etablerer ein sjølvstendig driftsorganisasjon, til dømes eit aksjeselskap, som kan organisere innsamlingsarbeidet og leggje til rette for distribusjon og andre tenester frå språkbanken. Alternativt kan desse oppgåvene gjerast av eksisterande verksemder. Hovudregelen bør vere at innsamlingsarbeidet blir sett bort som oppdrag, gjerne etter anbodsprinsippet, til eksisterande miljø med tilstrekkeleg kompetanse og med erfaring frå innsamling og distribusjon av språkdata. Senter for humanistisk informasjonsteknologi (HIT-senteret) er nær knytt til Universitetet i Bergen, og eit døme på eit slikt miljø. Eit anna er det språkteknologiske utviklingsmiljøet rundt Nordisk Språkteknologi (NST) og Sail Port Northern Europe as (SPNE) på Voss, men òg andre fagmiljø ved universiteta, og utanfor, kan tildelast innsamlingsoppdrag. Denne modellen vil òg vere kostnadssparande, då fleire av desse institusjonane er villige til å setje inn eit visst mon av eigne ressursar i arbeidet. Ein kan tenkje seg at det blir etablert ei form for rammeavtalar med institusjonar og miljø som har kompetanse til å ta på seg innsamlingsarbeid. Dei einskilde oppdraga kan så forhandlast nærare, eller – i somme høve – bli sette ut på anbod.

Styret må syte for at driftsorganisasjonen har klare retningsliner for korleis han skal følgje opp styret sine prioriteringar. Styret har ansvar for dei overordna prioriteringane mellom ulike typar språkressursar og tempoplan/progresjon i innsamlingsarbeidet. Dette kan gjerast gjennom rammeløyvingar med klare retningsliner, men likevel slik at driftsorganisasjonen har fleksibilitet til å gjere gunstige ad hoc-avtalar når det gjeld til dømes situasjonsbundne frikjøp av betydelege ressursvolum.

Når det gjeld oppgåver med å halde ved like og vidareutvikle språkressursane, kan dei implementerast på liknande vis som for innsamlingsarbeidet ved at oppgåvene blir behandla av språkbanken sin driftsorganisasjon. Her kan det likevel vere meir aktuelt med eit sett av samarbeidsavtalar med faginstitusjonar som tek seg av desse oppgåvene. Ein kan til dømes tenkje seg at universiteta får øyremerkte løyvingar over rammebudsjetta sine knytte til oppgåver med vedlikehald og vidareforedling av språkressursar. Kva for oppgåver som skal prioriterast, blir òg her bestemt av styret for språkbanken. Ei slik løysing vil sikre at vedlikehaldet blir utført av fagleg kompetente miljø. Denne modellen vil i tillegg ha den fordelen at det vil styrkje universitetsmiljøa, noko som vil gje meir forsking og auka

kompetanse. Samtidig gjev det grunnlag for å auke utdanningskapasiteten for høgt kvalifisert personell, og det vil kome industrien til gode.

Skjematisk oppsummering av korleis prosjektgruppa tenkjer seg organisasjonen for språkbanken:



Figur 2: Oganisasjonsstruktur

Prosjektgruppa ser for seg at stiftinga Norsk språkbank har ein liten, fast administrasjon, som har ein administrativ leiar på heiltid. I tillegg trengst det administrative ressursar til støtte for arbeidet med prioritering og planar, og for å definere og inngå avtalar og kontraktar. Desse ressursane kan ein tenkje seg stilt til rådvelde gjennom ein avtale med til dømes Norsk språkråd. Driftsorganisasjonen (driftsselskapet) vil få meir omfattande oppgåver, ettersom det er herfrå dei konkrete oppdraga skal setjast ut og følgjast opp. I stiftinga (rådet, styret og administrasjonen) og særleg i driftsselskapet er språkteknologisk kompetanse eit avgjerande krav, medan driftsselskapet òg må ha kompetanse og erfaring innanfor prosjektadministrasjon, oppfølging og gjennomføring innanfor fastlagde finansielle rammer.

KAPITTEL 4 INNHALD I EIN NORSK SPRÅKBANK

Innhaldet i ein nasjonal språkbank skal dekkje dei grunnleggjande behova til språkteknologisk forsking og produktutvikling, og vere ei hjelp til å redusere kostnadene ved utvikling og tilpassing av norske språkteknologiske produkt og tenester. Alle spørsmål som gjeld bruksrettar, må vere avklarte.

4.1 Innleiing

Innhaldet i språkbanken må oppfylle desse overordna krava:

- Materialet skal vere tenleg for norsk språkteknologisk industri og forsking
- Materialet må vere representativt for ulike språklege bruksområde (tekst, tale, dialektar, målformer), og i stor grad spegle av språkbruken i dag
- Materialet må vere godt dokumentert
- Koding og oppmerking av materialet må følgje internasjonale standardar
- Alle bruksrettar må vere avklarte

I rapporten "Norsk språkbank. Utredning om et nasjonalt korpus for språkteknologi" (Svendsen 1999) blei innhaldet i ei norsk samling av språkressursar delt inn i grove kategoriar som taledata (lydopptak av ymse slag), tekstdata (ulike tekstsamlingar) og leksikalske ressursar (ordlister, termbasar). Vi vil framleis halde oss til denne grovinndelinga.

I rapporten frå 1999 finn ein mange diskusjonar som er relevante for dette arbeidet og som prosjektgruppa har bygd vidare på i sine vurderingar. Dette gjeld til dømes grunngjevinga for val av ressurstypar. Vi har ikkje funne grunn til å ta dei opp att her, men for den som er interessert i fleire detaljar kring dette, viser vi til den rapporten.

I Nederland og Belgia er det nyleg gjort eit større arbeid med å spesifisere kva som bør inngå av data og grunnleggjande teknologimodular i ei generell ressurssamling for språkteknologi for nederlandsk. Spesifikasjonane kan generaliserast til andre språk, og er mellom anna nytta som grunnlagsmateriale for utforminga av ei innsamling av nye (ferske) språkressursar for fransk. Når det gjeld behovet for data, er det stor likskap mellom dei tilrådingane som blir gjorde her, og tilrådingane i den norske rapporten frå 1999.

Innhaldet i ein nasjonal språkbank skal dekkje dei grunnleggjande behova til språkteknologisk forsking og utvikling, og vere ei hjelp til å redusere kostnadene ved utvikling og tilpassing av språkteknologiske produkt og tenester for norsk. Prosjektgruppa har vurdert forslaga frå 1999 opp mot det vi meiner er behova i dag og i tida framover. Gruppa meiner at hovudkomponentane i det førre framlegget kan stå som dei er, men at nokre modifiseringar og tillegg bør gjerast.

Overslaga over kva som trengst, er høgare enn i Svendsen 1999. Det skuldast følgjande ekstra parametrar:

- Mykje meir av spontan tale - dette trekkjer mest opp på taledatadelen
- Det er lagt inn fleirspråklege tekstar for å støtte arbeid med maskinomsetjing
- Det er lagt meir vekt på arbeid med å samordne leksikalske data - mange kjelder krev stor innsats med omsyn til harmonisering

- Det er laga ein plan for å etablere databasar for omgrepsbeskrivingar og semantiske nettverk for norsk

4.2 Typar språkdata:

Ressurssamlinga har tre hovudkomponentar:

- Taledata (opptak av tale lagra elektronisk)
- Tekstdata (samlingar av tekst med eller utan merking)
- Leksikalske data (ordsamlingar som er generelle eller retta mot einskilde fagområde, til dømes terminologilister)

Dette samsvarer med rapporten frå Svendsen 1999 og det ein har samla inn for andre språk.

Taledata blir brukt til talegjenkjenning (tale-til-tekst, tale-til-lydskrift, tale-til-konsept) og talesyntese (maskinprodusert tale). Her deler ein mellom telefonidata som er tekne opp over telefonnettet med den kvaliteten og dei støykjeldene som finst der, og opptak frå kontor med den kvaliteten og dei støykjeldene som finst der. Desse data blir i liten grad brukte om kvarandre.

Tekstdata er nødvendige både for å kunne lage språkmodellar til til dømes talegjenkjenning og for å analysere korleis språket faktisk blir brukt. Dei fleste språkteknologiske bruksområda, til dømes talegjenkjenning, stavekontroll, grammatikkontroll og omsetjingsprogram, treng store tekstmengder for å fungere tilfredstillande.

Gode leksikalske data er ein nødvendig føresetnad for alle språkteknologiske bruksområde. Det finst ein god del leksikalske ressursar for norsk som er samla inn med offentlege midlar, der midlane har vore i form av grunntildelingar over statsbudsjettet eller via Noregs forskingsråd. Summen av desse ressursane utgjer eit omfattande materiale.

4.3 Prinsipp for prioritering

Målet med å lage ei samling av språkressursar er å

- medverke til å oppfylle målsetjinga om at norsk - i tale og skrift - skal vere det dominerande bruksspråket i Noreg
- frigjere verdiskapingspotensialet for språkteknologi, ikkje minst i offentleg sektor
- gjere det mogleg å drive språkteknologisk forsking der norsk språk er den sentrale empiriske komponenten
- stimulere norsk IKT-industri til å satse på språkteknologiske løysingar slik at ein ikkje sakkar akterut i høve til utanlandsk industri på området

Når det gjeld prioriteringar mellom dei ressurstypane som skal inn i samlinga, bør ein leggje vekt på følgjande prinsipp:

- Ressurstypen skal vere relevant for viktige bruksområde
- Språkressursen er pr i dag ikkje-eksisterande, ikkje tilgjengeleg eller har for dårleg kvalitet
- Ressursen skal kunne samlast inn og tilretteleggjast for språkbanken i løpet av eit avgrensa tidsrom

- Det er konkret etterspurnad etter ressurstypen frå språkteknologisk industri og/eller forsking
- Ressursen er viktig for strategiske forskingsprogram som er sette i gang

Det overordna prinsippet for prioritering er å syte for at ein så raskt og kostnadseffektivt som mogleg kan nå måla ovanfor.

Det er sett i gang kommersielle aktivitetar og aktivitetar innanfor forsking i Noreg som klart er relevante for å kunne realisere måla ovanfor:

- diktering av elektroniske journalar i helsesektoren
- automatiske opplysningstenester over telefon
- korrektur-, grammatikkontroll- og omsetjingsprogram
- hjelpeverktøy for funksjonshemma og personar med lese- og skrivevanskar

Ein gryande kommersiell aktivitet er i dag i gang i norske bedrifter. Men tilgang på relevante språkdata for desse aktivitetane er monaleg mindre enn det som finst for engelsk, og også mindre enn det som finst for svensk og dansk.

Sidan oppbygginga av språkbanken skjer samtidig med at Noregs forskingsråd har sett i gang eit langsiktig kompetanseoppbyggjande språkteknologiprogram, kalla Kunnskapsutvikling for norsk språkteknologi (KUNSTI), er det naturleg å relatere delar av prioriteringane i språkbanken til det denne satsinga treng. I forarbeida frå Noregs forskingsråd i samband med KUNSTI er det ein føresetnad at KUNSTI skal kunne basere sine prosjekt på at språkdata er tilgjengelege, noko som ikkje er tilfelle når det gjeld taledata, tekstdata eller leksikalske data. Prioriteringane når det gjeld innhaldet i ressurssamlinga, bør stette behov frå sentrale forskingsprosjekt som til dømes KUNSTI-programmet.

Ressurssamlinga må vidare prioritere slik at gode produkt som eksisterer for andre språk, innan rimeleg tid også kan kome på norsk. Kva som bør inngå i klassifikasjonen av denne typen produkt, bør industrien, brukargruppene og forskarsamfunnet samla kunne levere eit grunnlag for, og styret for språkbanken må gjere dei konkrete prioriteringane. Nordisk språkteknologi sitt arbeid med dikteringssystem, mellom anna i helsevesenet, er ein naturleg kandidat. Maskinomsetjing er ein annan mangel i norsk språkteknologisk produktutvikling. Det er naturleg å ha eit sideblikk til korleis andre land prioriterer innsamling av sine språkressursar. Initiativet frå Nederland og Belgia er svært interessant i denne samanhengen.

4.4 MINSTEINNHALD OG ØNSKT INNHALD

Prosjektgruppa har valt å konsentrere arbeidet om det minimumet av innhald som må vere i basen for at han skal bli brukande for målgruppene. Ønskt innhald er større, men vanskelegare å kalkulere, rett og slett fordi ein aldri får nok data til opptrening av statistisk baserte språkteknologiprodukt. Generelt kan ein seie at innhaldet i språkbanken bør vere godt over det minimum gruppa tilrår, og i oppsetta nedanfor er dei jamt over dobla.

Framlegga nedanfor er ikkje ”absolutte”. Styret for språkbanken må sjølvsagt kunne justere innhaldet dersom behova endrar seg.

4.4.1 Tale

Taledata er kjernen i all teknologi som gjeld gjenkjenning av tale og produksjon av tale (syntetisk eller kunstig tale). Talegjenkjenning krev opptak av mange talarar i ulike aldersgrupper og med ulike dialektar. Opptaka bør vere knytte til realistiske brukssituasjonar i den grad det lèt seg gjere. Dette gjeld til dømes talegjenkjenning i bil og mobiltelefon, gjenkjenning av spontan tale, diktering frå helsepersonell, advokatar osv. Slike opptak er kostbare å samle inn. Ofte må utviklarane bruke data som er samla inn i kontrollerte situasjonar, for å trene opp ein førstegenerasjonsapplikasjon. Når denne er laga, kan ein bruke han (til dømes til ein ruteopplysingstelefon) til å samle inn meir data (dette føreset at ein gjev den informasjonen Datatilsynet krev i slike høve, jamfør òg den juridiske utgreiinga).

For å utvikle ein applikasjon for kunstig tale trengst det ein einskild talar som les inn store mengder tekst med naturleg tonefall (prosodi) basert på eit gjennomtenkt vokabular. Vokabularet skal i den grad det er mogleg, dekkje ord og frasar som ein ventar å finne att i tekstar som seinare skal lesast opp av maskinen.

Eit minimumsomfang er ei samling med digitaliserte taleopptak tilsvarande 1700 timar tale (nær 17 millionar ord), der ein del er opplesing og resten spontan tale. Manuskriptlesen tale vil utgjere grunnlaget. Her vil ein kunne få den variasjonen i materialet som trengst for den grunnleggjande taleteknologien. Representasjon av ulike typar støy er òg viktig, men det kan til dels kompenserast ved simuleringar. Spontan tale er den taletypen som vil vere dominerande til bruk i taleteknologien. Rik representasjon av slik tale er av den grunn viktig. Prosodisk merking av naturleg, spontan tale er svært nyttig for å betre kvaliteten på syntetisk tale og for kvaliteten på neste generasjon av talegjenkjenning. Ei stor mengd transkribert spontan tale vil vere eit godt grunnlag for betre modellering av strukturelle fenomen i slik tale. Tilsvarande er tale frå dialogar mellom menneske viktig som grunnlag for opptrening av talegjenkjennarar til dialogsystem.

Om lag 900 timar skal vere manuskriptlesen tale, som skal danne grunnlag for akustiske modellar for talegjenkjenning. Fordelinga skal gjerast slik at begge målformer blir likestilt når det gjeld kvaliteten på dei taleteknologiske modellane. Denne delen vil òg inkludere taledata for utvikling og eksperimentering med syntetisk tale. Resten bør bestå av ulike typar spontan tale: diktering, menneske-maskin-dialogar, menneske-menneske-dialogar og samtale mellom fleire menneske. Materialet må vere fordelt mellom høgkvalitetsopptak og telefon- og mobiltelefonopptak. Det er avgjerande for bruksverdien å ha ei representativ dekning av dialektar, aldersgrupper, sosiolektar og kjønn. Desse opptaka skal som minimum vere merkte (annoterte) på ortografisk nivå, medan ein mindre del må merkjast på eit detaljert fonetisk og lingvistisk nivå.

I tillegg er det ønskjeleg med samlingar som inneheld ein vesentleg komponent tale:

- multimodale korpus, dvs. databasar som inneheld tale og data frå andre modalitetar som peiking, nikking, tastetrykk osv.
- fleirspråklege taledatabasar som kan nyttast til å finne samanhengar mellom ulike talte språk
- multimediale korpus som i tillegg til tale frå radio og fjernsyn har informasjon frå andre medium som tekstar og figurar frå verdsvev, aviser, tidsskrift osv.

Her har vi valt å prioritere data som er nødvendige for teknologiutvikling som er i gang eller planlagd sett i gang.

Tabell 4.1: Taledata, behov (modifisert etter Svendsen 1999):

			Minimum			Ønskt		
Type	Talestil	Formål	Timar, minumum	Årsverk forskar	Årsverk andre	Timar, ønskt	Årsverk forskar	Årsverk andre
Romkvalitet	Spontan	Diktering, dialogar	500	6,25	18,75	1000	12,50	37,50
Romkvalitet	Manuskript	Diktering, modellar	500	4,17	12,50	1000	8,33	25,00
Telefon	Manuskript	Modellar	120	0,60	1,80	240	1,20	3,60
Mobiltlf.	Manuskript	Modellar	120	0,75	2,25	240	1,50	4,50
Telefon i bil	Manuskript	Diverse	120	3,00	9,00	240	6,00	18,00
Telefon	Spontan	Dialogar	100	1,25	3,75	200	2,50	7,50
Telefon	Spontan	Diktering	100	1,25	3,75	200	2,50	7,50
Romkvalitet	Manuskript	Difondatabase	2	1,00		4	2,00	
Romkvalitet	Manuskript	Prosodi / lydbibliotek	20	1,00	0,50	40	2,00	1,00
Telefon	Manuskript	Emnesøk i multimediearkiver	20	1,00	0,50	40	2,00	1,00
Audio	Spontan	Emnesøk	100	1,25	3,75	200	2,50	7,50
Romkvalitet	Spontan	Multimodale grensesnitt				100	1,25	3,75
Romkvalitet	Spontan	Modellar, fleirspråklege applikasjonar				100	1,25	3,75
Høg romkvalitet	Manuskript	Konkatenativ talesyntese	40	0,25	0,75	80	0,50	1,50
SUM			**1742**	**21,77**	**57,30**	**3684**	**46,03**	**122,1**

4.4.2 Tekst

Den største delen av tekstmaterialet skal vere norske tekstar som er automatisk merkte med omsyn til ordklasse – minst om lag 100 millionar ord for kvar målform. Tekstane bør vere delte mellom sakprosa, småtrykk, upublisert materiale, aviser og andre trykte media, dessutan skjønnlitteratur. Ein liten del skal vere særleg lag til rette for opptrening og validering av statistiske språkanalyseprogram som gjer ein nokså overflatenær (shallow) analyse av tekst.

Omfattande tekstdatabasar er ei primærkjelde for utvikling av leksikalske ressursar og statistiske språkmodellar for talegjenkjenning. Utforminga av tekstsamlingane må ta omsyn til dette. Forslaget til omfang av tekstbasen er eit absolutt minimum.

Tekstbasane må vere merkt etter tilrådingane frå *Text Encoding Initiative* (TEI). Dette er i samsvar med tilrådinga frå Svendsen 1999.

I tillegg trengst det fleirspråklege parallellkorpus. Dette er heilt avgjerande tekstsamlingar for maskinomsetjing, anten det gjeld omsetjing mellom bokmål og nynorsk eller mellom norsk og eit framandspråk. Parallellkorpus er òg informasjonskjelder for konstruksjon av semantisk strukturerte leksikalske databasar, som er verdfulle for til dømes informasjonssøking og tekstsamandrag. I det minste bør eit parallellkorpus for norsk-engelsk inkluderast i den norske språkbanken, helst bør fleire språkpar vere representerte. Det meste skal leggjast til rette slik at original og omsetjing blir kopla saman setning for setning, og ein mindre del ord for ord.

Utover dette trengst det såkalla "trebankar" der ein lagrar korrekte strukturar for setningar. Slikt materiale er nødvendig mellom anna i utviklinga av statistiske modellar for syntaksanalyse, til dømes som treningsgrunnlag for program som sjølve skal lære grammatiske strukturar frå tekst og nytte kunnskapen i setningsanalyse. Denne ”sjølvlærande” tilnærmingsmåten er særs viktig for å lage analyseprogram med stor dekningsgrad.

Det er også teke med ei samling treningsdata særskilt retta mot medisinsk diktering. På dette området er effektiviseringsgevinsten i offentleg sektor stor.

Tabell 4.2: Tekstdata, behov

		Minimum			Ønskt		
Teksttypar	Tilarbeiding	Omfang	Forskar	Andre	Omfang	Forskar	Andre
Tospråklege tekstar (no-en)	Basal tilrettelegging	2 500 000	0,71	2,14	5 000 000	1,43	4,29
Tospråklege tekstar (en-no)	Basal tilrettelegging	2 500 000	0,71	2,14	5 000 000	1,43	4,29
Tospråklege tekstar (en-no og no-en)	Grundig tilrettelegging	500 000	0,71	2,14	1 000 000	1,43	4,29
Sakprosa, små-trykk, upublisert materiale	Basal tilrettelegging	50 000 000	1,43	4,29	100 000 000	2,86	8,57
Aviser og medium, skjønnlitteratur	Basal tilrettelegging	50 000 000	1,43	4,29	100 000 000	2,86	8,57
Sakprosa, små-trykk, upublisert materiale	Utvida tekstkoding, manuelt kontrollert ordklassemerking	500 000	0,24	0,71	1 000 000	0,48	1,43
Aviser og medium, skjønnlitteratur	Utvida tekstkoding, manuelt kontrollert ordklassemerking	500 000	0,24	0,71	1 000 000	0,48	1,43
Aviser	Etablering av trebank	200 000	0,29	0,86	1 000 000	1,43	4,29
Anonymiserte journalar	Treningsdata for medisinsk diktering	0			200 000	1,00	3,00
			11,52	34,57		26,76	80,29

4.4.3 Leksikalske ressursar

Denne delen av språkbanken vil innehalde leksikon og tesaurusar. Med leksikon meiner ein her elektroniske ordlister med informasjon om ordforrådet i eit språk på ulike språkvitskaplege nivå. Tesaurusar er leksikon med semantiske og assosiative relasjonar mellom ord. Det kan inkludere emnetesaurusar (til dømes for fagterminologi).

Dei leksikalske ressursane vil vere grunnordlister (allmennord, termar, namn osv.) med tilgang til fullformsproduksjon (dvs. korleis orda er bøygde i bøyingsmønster, jamfør til dømes ressurs – ressursen – ressursar – ressursane), uttaleleksikon, dialekt- og rettskrivingstesaurusar og emnetesaurusar (synonymordlister og fagspesifikke semantiske ordlister).

Det finst alt tilgjengeleg omfattande ordlister for norsk (bokmål og nynorsk) som er utvikla ved universiteta. Desse har grunnformer, bøyingsprogram, lydbeskrivingar og fullformslister. Nordisk Språkteknologi har samla inn eit materiale på totalt 1,5 millionar ord.

Det er altså samla inn mykje materiale som kan gjenbrukast. Minstebehovet for leksikalske data (ordlister) er 500 000 ordformer pr målform. Alt materiale som blir lagt inn i språkbanken, må kvalitetssjekkast og standardiserast når det gjeld grammatisk informasjon,

rettskrivingskonvensjonar og uttalestandard. Det må vidare lagast meir djuptgåande spesifikasjonar av uttale av namn, framandord og nyord, og uttalespesifikasjonane må tilpassast til dei ulike dialektområda.

Industri og forsking etterspør ofte ein norsk versjon av det engelske "Wordnet", som har eksistert i om lag ti år. Denne typen ressursar kan mellom anna nyttast i informasjonsgjenfinning og omsetjingsprogram, og vi har derfor teke med dette som ein ressurs språkbanken bør kunne tilby. Til slutt har vi inkludert ein omgrepsspesifikasjon der ein koplar ein omgrepdatabase for norsk til ein tilsvarande engelsk. Dette er ein norsk variant av eksisterande EU-ressursar.

Tabell 4.3: Leksikalske ressursar, behov

	Minsteomfang				Ønskt omfang			
Aktivitetstype	Tal på ord	Forskar	Andre	Innkjøps-kostnader	Tal på ord (fullformer)	Forskar	Andre	Innkjøps-kostnader
Ordlistedata, bm *)	500 000			1 666 667	1 000 000			3 333 333
Ordlistedata, nn *)	500 000			1 666 667	1 000 000			3 333 333
Ordlister frå ulike kjelder		2,00	1,00			2,00	1,00	
Utvikling av stavevariantar /basis dialektvariantar		1,00	1,00			2,00	2,00	
Utvikling av uttalespesifikasjon for navn, framandord, nyord		1,00	3,00			0,50	3,00	
Kvalitetskontroll av eksister-ande lister (bm og nn)		2,00				2,00		
Uttalespesifikasjon for dialektregionar		1,00	3,00			1,00	4,00	
Ordnett (norsk Wordnett)	50000	0,71	2,14		100000	1,00	3,00	
Omgrepstydingar – SIMPLE	50000	0,71	2,14		100000	1,00	3,00	
SUM		8,42	12,28	3 333 334		9,50	16,00	6 666 666

4.5 Standardar

Materialet som blir samla inn, må i størst mogleg grad vere tilpassa EU-standardar som Expert Advisory Group for Language Engineering Standards (EAGLES) og TEI. Når det gjeld standardar og bruk av desse i materiale som blir samla inn for språkbanken, er det eit ufråvikeleg krav at internasjonale standardar blir brukte så langt dei finst. I fall det ikkje finst standardar, skal innsamlarane retta seg etter det som blir brukt internasjonalt (jamfør ELDA/ELRA).

4.6 Verktøy

Verktøy som er utvikla eller skaffa i samband med innsamling og foredling av data for språkbanken, må inngå som ressursar i språkbanken og stillast til rådvelde for andre. Det kan gjelde til dømes programvare for innlesing og opptak av tale, transkribering og annotering, analyse, konvertering mellom ulike dataformat osv.

Taggarane utvikla ved Universitetet i Oslo i samarbeid med HIT-senteret i Bergen til automatisk merking av tekstar, vil vere tilgjengelege for ressurssamlinga. Desse taggarane har eit presisjonsnivå på om lag 95 %, som er tilfredsstillande for automatisk merking av tekst.

KAPITTEL 5 KOSTNADSOVERSLAG

5.1 Innleiing

Hovuddelen av kostnadene til etablering av dei basale ressursane som bør gå inn i ein norsk språkbank, er knytt til arbeidsomfang. Det er mogleg å kjøpe rettar til eksisterande datasamlingar, men i ein del tilfelle vil kanskje kostnadene ved kjøp ikkje liggje så langt under kostnadene med ei ny innsamling. Av den grunn har vi valt å la arbeidsomfanget liggje til grunn for kostnadsestimatet. Dermed oppnår vi også ei verdifastsetjing av eksisterande materiale. Prosjektgruppa har heller ikkje teke standpunkt til kor mykje av det eksisterande materialet som kan nyttast, og kor mykje som må samlast inn på nytt for å vere brukande i språkbanken. Dette må bli ei oppgåve for språkbanken sitt strategiske organ.

Arbeidsomfanget er så langt råd er, brote ned på årsverk, som igjen er knytt til arbeidsomfang per time for tekstdata og leksikalske data, medan det er kvantifisert som timar tale per årsverk for taledata. I kostnadsestimata for datainnsamlinga er følgjande lagt til grunn

- 25 % av alt arbeid må vere leidt av ein forskar eller andre med solid erfaring og kompetanse på området
- Validering av data er inkludert i modellen
- SND sitt kostnadsoverslag for høgt kvalifisert arbeidskraft og konsulentar på lågare nivå ligg til grunn for kalkylane over personalkostnadene
- Det er ein føresetnad at utstyr og relevante verktøy er dekte innanfor rammene av arbeidstimane

Det har vore vanskeleg å innhente nøyaktige estimat. Estimata i denne rapporten byggjer på den førre rapporten, erfaringar blant medlemmene av prosjektgruppa og samanlikningar med tilsvarande prosjekt i andre land. Det er eit problem at få ønskjer å bli siterte på kva som er eksakt omfang. Ein må resonnere seg fram til brukelege estimat. Estimata her er i dei fleste tilfella basert på erfaringar som medlemmer i prosjektgruppa har gjort, dvs. innsamlingar ved NST, Telenor, Universitetet i Oslo og NTNU.

Tabell 5.1: Parametertal for berekning av kostnader

Lønn forskar (parametertal)	800 000
Lønn assistentar (parametertal)	500 000
Tal på timar, manuslesen tale per årsverk	30
Tal på timar, spontan tale per årsverk	20
Tal på ord behandla per time for tekstkorpus	5 000
Tal på ord behandla per time for bilingvale korpus, ein vei	500
Tal på ord behandla per time manuelt kontrollerte POS-korpus	300
Tal på ord nøyaktig (ord-ord) samankopling bilingvale korpus	100
Tal på timar opptak fasttelefon per årsverk	50
Tal på timar opptak mobiltelefon per årsverk	40
Tal på timar opptak i bil per årsverk	10
Tal på timar innlesen konkatenativ talesyntese per årsverk	40

Leksikalske data – Tal på ord per time transkribert og grammatisk kontrollert	100
Trebankar – tal på ord per time	100
SIMPLE, Wordnet – tal på tydningseiningar per time (="ord/omgrep")	10

5.2 Taledata

Kostnadene til innsamling av taledelen for det nederlandske korpuset er rekna til 5 millionar euro, om lag 38 millionar kr. Dette korpuset er på 1000 timar tale, dvs. noko mindre enn det som blir tilrådd for norsk. Grunnen til at den norske talemålsdelen er større, er for det første omsynet til dekning av manuskriptlesen tale både på bokmål og nynorsk, og for det andre var det tidlegare samla inn ein del data i Nederland. Hovuddelen av arbeidet blir gjort av universitet i Nederland og Belgia, slik at kostnadsnivået blir halde på eit minimumsnivå.

I rapporten frå 1999 blei det gjort eit overslag over kostnadene for å samle inn dataressursane. Desse overslaga samsvarer godt med dei erfaringane NST har gjort ved innsamlinga av sitt eige materiale. Ein kan byggje på verktøy og erfaringar til dømes frå Nederland når ein skal til med innsamling her i landet, og det kan få arbeidsomfanget og dermed kostnadene per time talemateriale noko ned. Vi tek derfor utgangspunkt i kostnaden for det nederlandske talemålskorpuset. Taledelen av ein norsk språkbank blir estimert til eit arbeidsomfang på vel 57 årsverk.

Tabell 5.2: Taledata, kostnader

Type	Talestil	Formål	Kostnad, minimum	Kostnad, ønskt nivå
Romkvalitet	Spontan	Diktering, dialogar	14 375 000	28 750 000
Romkvalitet	Manuskript	Diktering, modellar	9 583 333	19 166 667
Telefon	Manuskript	Modellar	1 380 000	2 760 000
Mobiltlf.	Manuskript	Modellar	1 725 000	3 450 000
Tlf. i bil	Manuskript	Diverse	6 900 000	13 800 000
Telefon	Spontan	Dialog	2 875 000	5 750 000
Telefon	Spontan	Diktering	2 875 000	5 750 000
Romkvalitet	Manuskript	Difondatabase	800 000	1 600 000
Romkvalitet	Manuskript	Prosodi/lydbibliotek	1 050 000	2 100 000
Telefon	Manuskript	Emnesøk i multimediearkiv	1 050 000	2 100 000
Audio	Spontan	Emnesøk	2 875 000	5 750 000
Romkvalitet	Spontan	Multimodale grensesnitt	0	2 875 000
Romkvalitet	Spontan	Møtetranskripsjon	0	2 875 000
Høg romkvalitet	Manuskript	Konkatenativ talesyntese	575 000	1 150 000
SUM			**46 063 333**	**97 876 667**

5.3 Tekstdata

Arbeidsomfanget knytt til tekstdelen var førre gongen estimert til knappe 40 årsverk. Skal det òg leggjast inn eitt eller fleire parallellkorpus i samlinga, vil det auke omfanget noko, og estimatet er på 45 årsverk. Dette er det teke omsyn til i talet på årsverk i nedste rada. Merk at det ikkje er sett av ressursar til frikjøp eller kompensasjon for tekstar som blir tekne inn i

basen. Dersom ein må betale slik kompensasjon, vert kostnadane høgare. Sjå også avsnitt 7.13.

Tabell 5.3: Kostnader, tekstdata

Teksttypar	Tilarbeiding	Kostnad, minimum (bm & nn)	Kostnad, ønskt nivå (bm & nn)
Tospråklege tekstar (norsk - engelsk)	Basal tilrettelegging	3 285 714	6 571 429
Tospråklege tekstar (engelsk - norsk)	Basal tilrettelegging	3 285 714	6 571 429
Tospråklege tekstar (engelsk - norsk og norsk - engelsk)	Grundig tilrettelegging	3 285 714	6 571 429
Sakprosa, småtrykk, upublisert materiale	Basal tilrettelegging	6 571 429	13 142 857
Aviser og medium, skjønnlitteratur	Basal tilrettelegging	6 571 429	13 142 857
Sakprosa, småtrykk, upublisert materiale	Utvida tekstkoding, manuelt kontrollert ordklassemerking	1 095 238	2 190 476
Aviser og medium, skjønnlitteratur	Utvida tekstkoding, manuelt kontrollert ordklassemerking	1 095 238	2 190 476
Aviser	Etablering av trebank	1 314 286	6 571 429
Anonymiserte journalar	Treningsdata for medisinsk diktering	0	4 600 000
SUM		**26 504 762**	**61 552 382**

5.4 Leksikalske ressursar

I Svendsen 1999 tok omtalen av den leksikalske databasen utgangspunkt i at mykje av dei basale ressursane alt er på plass, både for bokmål og nynorsk. Det meste av arbeid var tenkt brukt til å leggje til variantar av staving, uttale og utarbeiding av uttalebeskrivingar for namn. Semantiske tesaurusar og emnespesifikke tesaurusar (som synonymordlister innanfor ulike fagdisiplinar) blei ikkje prioriterte. Fleirspråklege leksikon blei ikkje vurderte.

Ved førre krossvegen blei arbeidsomfanget til å etablere dei leksikalske ressursane estimert til 10 årsverk. Grunna nokre utvidingar av innhaldet er estimatet justert til 21 årsverk.

Tabell 5.4: Kostnader, leksikalske data

Aktivitetstype	Kostnad, minimum	Kostnad, ønskt nivå
Innkjøp ordlistedata, bokmål	1 666 667	3 333 333
Innkjøp ordlistedata, nynorsk	1 666 667	3 333 333
Innleming av ordlister frå ulike kjelder	2 100 000	2 100 000
Utvikling av stavevariantar/basale dialektvariantar	1 300 000	2 600 000
Utvikling av uttalespesifikasjon for namn, framandord og nyord	2 300 000	1 900 000
Kvalitetskontroll av eksisterande lister	1 600 000	1 600 000
Uttalebeskrivingar for dialektregionar	2 300 000	2 800 000
Ordnett	1 642 857	2 300 000
Omgrepsontologi (SIMPLE)	1 642 857	2 300 000
SUM	16 219 048	22 266 666

5.5 Administrative kostnader

Dei administrative kostnadene med innsamlingsarbeidet byggjer på erfaringar frå NST og SPNE, dessutan meir generelle erfaringar frå anna distribusjonsarbeid. Kostnadene er høgast det første året då innsamlingsarbeidet skal planleggjast og anbodsutlysingar skal lagast. Så kan ein rekne med at kostnadene går ned til ca. 1,5 millionar kr per år.

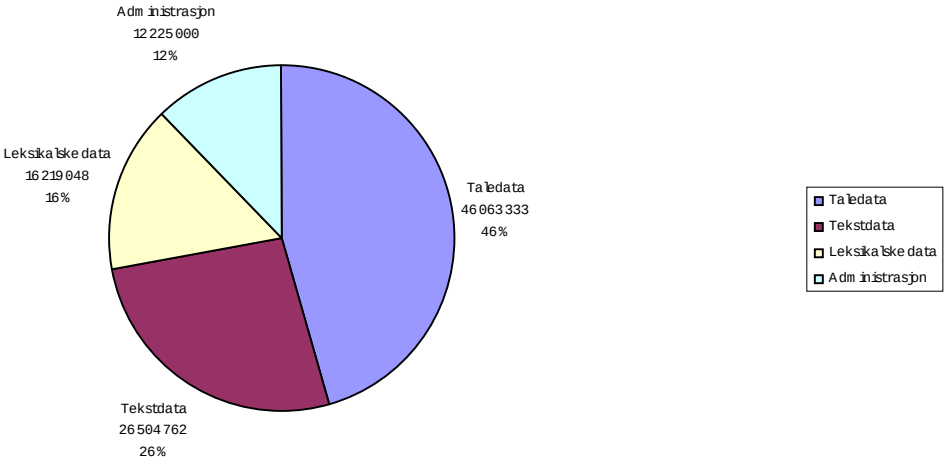
Tabell 5.5: Administrative kostnader

Administrasjonskostnader											
	År 1		**År 2**		**År 3**		**År 4**		**År 5**		
Oppgåve	**Årsv.**	**Kostnad**	**Årsv.**	**Kostnad**	**Årsv.**	**Kostnad**	**Årsv.**	**Kostnad**	**Årsv.**	**Kostnad**	**Totalt**
Adm.leiar, stifting	1,0	800 000	1,0	800 000	1,0	800 000	1,0	800 000	1,0	800 000	4 000 000
Driftskostn., stifting, infrastruktur		250 000		250 000		150 000		150 000		150 000	950 000
Internasj. kont., stifting		75 000		75 000		75 000		75 000		75 000	375 000
Jur. hjelp, stifting		200 000		50 000		50 000		50 000		50 000	400 000
Adm.leiar, driftsorg.	1,0	800 000	1,0	800 000	1,0	800 000	1,0	800 000	1,0	800 000	4 000 000
Konsulent, driftsorg.	1,0	500 000	1,0	500 000	1,0	500 000	1,0	500 000	1,0	500 000	2 500 000
SUM		**2 625 000**		**2 475 000**		**2 375 000**		**2 375 000**		**2 375 000**	**12 225 000**

5.6 Totalkostnad

Det samla omfanget med innsamling og tilrettelegging av data vil vere rundt 150 årsverk. Ein kan gå ut frå at opp mot ein tredel av dette vil dreie seg om høgt kvalifiserte personar, medan resten kan utførast av "ufaglærde". Utgangspunkt er i ein gjennomsnittleg kostnad på 800 000 kr/årsverk for dei høgast kvalifiserte og 500 000 kr/årsverk for dei andre. 150 årsverk vil gje ein kostnad på rundt 90 millionar kr. I tillegg kjem kostnadene til administrasjon, infrastruktur, godtgjersle til informantar, reiser osv. på rundt 10 millionar kr. Den totale etableringskostnaden for ein norsk språkbank vil vere på rundt 100 millionar kr.

Visuelt kan ein framstille fordelinga av kostnadene mellom dei ulike kategoriane slik:



Figur 3: Kostnader fordelt på kategoriar

KAPITTEL 6 FINANSIERING

Realistisk sett må finansieringa vere offentleg. Norsk språkteknologisk industri har ikkje styrke til å kunne vere med på finansieringa i særleg grad. Større land og språksamfunn i Europa enn Noreg finansierer hovuddelen av ressursane med offentlege midlar, trass i at dei har sterkare språkteknologisk industri og fleire ressurssamlingar frå før.

6.1 Innleiing

I prosjektgruppa sitt mandat inngår eit oppdrag om å utarbeide ein finansieringsplan. I mandatet er denne oppgåva definert slik:

> *"Denne (finansierings)planen må skissere hva det koster å etablere selve banken, hva det koster å skaffe og legge inn så mye ressurser at relevante brukergrupper kan nyttiggjøre seg innholdet (kritisk suksessfaktor), en opptrappingsplan for komplett samling, og hva det koster å drifte og utvikle selve språkbanken. Økonomien baseres på at dette blir en kombinasjon av privat og offentlig finansiering. Finansieringskilder identifiseres. Tiltakets sektorovergripende karakter skal gjenspeiles i de finansieringsmodeller som foreslås."*

Kostnadssida er behandla i kapittel 5.

I dette kapitlet vil vi drøfte prinsipp, føresetnader og moglege finansieringskjelder for ein norsk språkbank. Vekta er lagd på ein diskusjon om kva for prinsipp og føresetnader ein bør leggje til grunn for å finansiere etableringsfasen, dvs. nødvendig innsamlings- og tilretteleggingsarbeid for å etablere ei ressurssamling ut frå dei tilrådde minstekrava. I tillegg vurderer vi forhold knytte til finansiering av vedlikehald, drift og vidareutvikling av ressurssamlinga.

6.2 Føresetnader og prinsipp for finansiering

Å byggje opp ei norsk ressurssamling for språkteknologi er å etablere ein nasjonal infrastruktur som tener kulturelle, samfunnsøkonomiske og næringsmessige formål. Døme frå andre europeiske land tilseier at ein slik infrastruktur primært bør og må vere offentleg finansiert – det dreier seg om å samle inn og forvalte ressursar som representerer ein felles nasjonal eigedom med klare samfunnsøkonomiske gevinstar, gjennom ei samla forvalting av desse.

Sidan talet på medlemmer i det norske språksamfunnet er lågt samanlikna med andre land i Europa, er det av særleg stor kulturpolitisk verdi å ha ei samling av norske språkteknologiressursar. Dette betyr at marknaden er så liten i Noreg at det ikkje lønner seg eller er økonomisk forsvarleg for enkeltaktørar, anten dei er i privat eller offentleg sektor, å finansiere innsamling og tilrettelegging av språkressursar for anna enn heilt spesifikke og svært avgrensa formål.

Utgreiingsarbeidet viser at dette er eit område med klare stordriftsfordelar – språkteknologiske ressurssamlingar må ha om lag same omfang uavhengig av språket si utbreiing og talet på brukarar. Det er desse vurderingane som har ført prosjektgruppa fram til konklusjonen om at etableringa av ein norsk språkbank er avhengig av at grunninvesteringane blir dekte gjennom særskilt finansiering over offentlege budsjett. Føresetnaden blir forsterka av at Noreg er relativt seint ute med planar for å byggje opp ei nasjonal samling av språkressursar. Det hastar med å setje i verk planen. Å basere gjennomføringa på privat finansiering er urealistisk og vil berre føre til ytterlegare forseinkingar.

Grunngjevinga for å etablere ein norsk språkbank tilseier at hovudansvaret for å finansiere dei nødvendige basisinvesteringane ligg hos tre departement: Kultur- og kyrkjedepartementet (KKD), Nærings- og handelsdepartementet (NHD) og Utdannings- og forskingsdepartementet (UFD). For KKD sin del er ei slik rolle tufta på den kulturpolitiske dimensjonen med det ansvaret dette departementet har for å skjøtte og styrkje norsk språk og kultur. NHD peiker seg ut som ein sentral aktør når det gjeld nyetablering og industriutvikling, dessutan har dette departementet samordningsansvaret for norsk IT-politikk. Språkteknologi er eit av fleire relevante område som er peikte på som viktige IT-politiske instrument, og som eit område der produktutvikling kan gje store effektiviseringsgevinstar, særleg i offentleg sektor. eNoreg-planen viser dette tydeleg, og planen ser språkteknologien som ein del av ein klar strategi for norsk digitalt innhald.

Utdannings- og forskingsdepartementet (UFD) er ein tredje sentral aktør når det gjeld grunninvesteringane, ut frå kor viktig ei slik ressurssamling vil vere som infrastruktur for språkvitskapeleg og språkteknologisk forsking.

Dei kulturpolitiske, IT- og næringspolitiske og forskingspolitiske dimensjonane ved ein slik infrastruktur som språkbanken er, skal avspeglast i den relative finansieringsfordelinga. Ei fordeling 3 : 3 : 2 mellom KKD, NHD og UFD er rimeleg. Det er viktig at dei tre departementa tek eit felles ansvar for å skaffe den nødvendige finansieringa. Språkbankens formål og den typen materiale som skal inngå, gjer det unaturleg og lite meiningsfylt å tilordne kvart einskilt av desse tre departementa finansieringsansvaret for ulike delar av samlinga.

Ut frå føresetnaden om at språkbanken blir etablert som ein sjølvstendig organisasjon, kan praktiske omsyn tale for at grunninvesteringane blir kanaliserte direkte frå dei involverte departementa til organisasjonen. Alternativt kan departementa kanalisere midlane som øyremerkte (ad hoc) gjennom underliggjande organ som blir bedne om å ta ansvar for vesentlege bruks- og brukarinteresser. Norsk språkråd (vis-à-vis KKD) og Noregs forskingsråd (vis-à-vis NHD og UFD) vil vere dei mest nærliggjande vala i så måte. Eit slik opplegg vil gje dei eit tydelegare fagleg og operativt medansvar i etableringa av språkbanken. For begge institusjonane vil etableringa av ein språkbank føre til at dei vil få føresetnader for og forventningar om å engasjere seg sterkare i oppgåver som kan dra nytte av norske språkteknologiske ressursar. Midlar til å etablere språkbanken må derfor gjevast særskilt, og ikkje som øyremerkte midlar på ordinære budsjett. Sistnemnde alternativ vil setje desse institusjonane sin sjanse til å finansiere den daglege verksemda under press, og dermed svekke evna deira til å vere aktive medspelarar for å få den beste bruken av infrastrukturen som blir bygd opp.

Dei fleste departementa er ansvarlege for verksemder som vil ha brukarinteresser i ein norsk språkbank ved at bruksområda for språkteknologi blir utvida og aktualiteten aukar når det gjeld effektivisering av offentleg forvaltning og modernisering av offentlege tenester. På same måte som for private aktørar er det ikkje rimeleg å rekne med at brukarinteressene i høve til ein nasjonal infrastruktur skal koplast med eit medansvar for å finansiere nødvendige investeringar for å etablere infrastrukturen som må til. Praktiske omsyn taler mot ein altfor kompleks finansieringsstruktur. Det vil skape ein uoversiktleg og lågare grad av tryggleik for at finansieringa er på plass når det gjeld grunnlagsinvesteringane. Det vil også kunne føre til unødvendig administrativt ekstraarbeid når det gjeld å gjennomføre etableringa av ressurssamlinga.

Sjølv om tilrådingane frå prosjektgruppa er at finansieringa av grunninvesteringane for språkbanken blir knytt til nokre få departementale finansieringskjelder, er det ein føresetnad at andre private og offentlege aktørar yter monalege bidrag til (basis)samlinga. Dette vil dels vere som hjelp og bidrag til faglege og operative oppgåver i oppbygging av samlinga, dels ved å stille innsamla språkressursar til disposisjon (på nærare avtalte vilkår), dels ved å inngå samarbeid med språkbanken ved eigeninitiert innsamling av språkressursar (prosjekt)spesifikke oppgåver. For å få ei oversiktleg organisering og heilskapleg prioritering i oppbygginga av språkbanken er det mest formålstenleg å basere seg på denne typen bidrag frå dei viktigaste eigar- og brukarinteressene. Realistisk sett er omfanget av direkte finansieringsbidrag til grunninvesteringar i språkbanken liten, og denne forma for bidrag vil lett føre til at arbeidet med oppbygginga av språkbanken blir adhocprega, lite planmessig og utsett for situasjonsbestemte prioriteringar. Gjennom å mobilisere sentrale brukarinteresser i oppbyggingsfasen vil ein få nyttige erfaringar med ein samarbeidsmodell for dagleg drift og vedlikehald av språkbanken.

Prosjektgruppa vurderer dei offentlege finansieringsinstitusjonar (SND, SIVA, Argentum, Noregs forskingsråd sitt apparat for næringsretta FoU) retta mot næringsutvikling som viktige støttespelarar når det gjeld finansiering av FoU-prosjekt baserte på bruk av den typen språkressursar som språkbanken skal innehalde. Prosjektgruppa vurderer ikkje desse institusjonane slik at dei kan eller skal spele noka naturleg rolle i å finansiere denne typen nasjonal infrastruktur. Deira rolle og engasjement bør i denne samanhengen heller vere å styrkje finansieringsgrunnlaget for næringsverksemd retta mot realisering av norskspråklege produkt og tenester basert på språkteknologi.

6.3 Finansieringsløysingar

Kostnadsoverslaga som ligg til grunn for finansieringsmodellen som er skissert ovanfor, vil krevje ei finansiering over statsbudsjettet på ca. 100 millionar kroner over ein femårsperiode, dvs. om lag 20 millionar årleg. Fordelinga mellom dei tre departementa er av underordna praktisk verdi, og den skisserte fordelingsnøkkelen må departementa gjerne forhandle om. For å gjennomføre oppbygginga av språkbanken effektivt må ein lage ein heilskapleg plan for den aktuelle oppbyggingsperioden. Det er avgjerande for oppbygginga at dei aktuelle departementa kan gå inn på den samla finansieringspakka som ligg i tilrådingane frå prosjektgruppa. Korleis departementa vurderer dette, og om det er mogleg å realisere, kan ikkje prosjektgruppa dømme om utover å meine at det tilrådde omfanget av basen er godt grunngjeve og underbygd.

Dersom ei offentleg hovudfinansiering kan sikrast, er det truleg at det vil danne eit godt grunnlag for samarbeid med dei viktigaste offentlege og private aktørane som kvar på sitt vis kan medverke til at oppbygginga av språkbanken kan gjennomførast på ein vellykka og kostnadseffektiv måte. Både i offentlege forskingsinstitusjonar og hos private aktørar finst det (større og mindre) språkressurssamlingar som det kan vere aktuelt å leggje inn i språkbanken. Mange av dei sentrale aktørane uttrykkjer stor vilje til samarbeid om dette og om framtidig innsamlingsarbeid. Dette er eit godt grunnlag for at ein kan etablere ei slik ressurssamling med vesentleg mindre kostnader enn om språkbanken må setje ut heile arbeidet på rein oppdragsbasis.

Nokre tankar om det moglege bidraget frå Universitetet i Oslo kan illustrere dette: Her har ein i mange år arbeidd med tilrettelegging av elektroniske tekstar. Den faglege ekspertisen finst, og det er interesse for å gå inn i arbeidet med ei nasjonal ressurssamling. Det finst også mykje

erfaring frå organisasjon og leiing av store prosjekt. Leiinga ved Det historisk-filosofiske fakultetet er positive til eit nasjonalt prosjekt, og meiner det bør være mogleg å frigjere personalressursar som delfinansiering om det blir aktuelt å leggje større delprosjekt til fakultetet.

Det same gjeld Universitetet i Bergen, der ein har arbeidd med språkteknologi heilt frå først i 1970-åra. Ved Det historisk-filosofisk fakultetet finst det ekspertise på toppnivå med lang erfaring frå organisering av oppdragsforsking og anna eksternfinansiert forsking og utvikling.

Noregs forskingsråd sette i år (2002) i gang eit språkteknologisk forskingsprogram (KUNSTI). Denne satsinga er basert på ei oppfatning om at språkteknologi vil få ei så stor og veksande rolle at ein norsk språkressurssamling ganske enkelt må etablerast. Programmet siktar mot å styrkje det nasjonale forskingsmiljøet slik at det auka potensialet for norsk språkteknologisk FoU som ein språkbank vil representere, kan bli best mogleg brukt. Gjennom si store kontaktflate til det norske fagmiljøet vil KUNSTI-programmet vere ein aktiv medspelar når det gjeld å realisere eit breitt nasjonalt samarbeid om så vel faglege som operative sider av prosjektet.

Etablering av ei norsk ressurssamling for språkteknologi er å etablere ein infrastruktur både for forsking og industri. I mange store europeiske land har det offentlege saman med EU-midlar kosta det aller meste av grunninvesteringane som har vore nødvendige for enklare språkteknologiske produkt, medan teknologiselskapa har teke utviklingskostnadene knytte til sjølve produktutviklinga. I nokre tilfelle har prioriteringane for kva for materiale som skal samlast inn først, blitt snudd på ved at industrien har skote inn pengar for å få det gjort.

Den norske språkteknologiindustrien er i dag svært liten samanlikna med internasjonale selskap som Nuance, SpeechWorks, språkteknologiavdelingane til Philips, IBM, Siemens osv. Nordisk Språkteknologi på Voss er den største norske aktøren. Ved Telenor FoU har ei gruppe forskarar arbeidd med taleteknologi i vel 20 år. I tillegg finst det nokre små firma ulike stader i landet. Ein kan ikkje rekne med at desse har økonomi til å finansiere særleg store mengder av det som trengst i ei slik samling som denne rapporten siktar mot. Dei vil derimot ha god bruk for innhaldet, som vil setje dei i stand til å utvikle norskspråklege produkt. I tillegg vil ressurssamlinga gjere det mogleg for norsk industri å lage nye produkt som i sin tur kan tilpassast andre språk. Her er det som i industriverksemd generelt viktig å kome i gang på ein heimemarknad før ein ekspanderer internasjonalt. Ressurssamlinga vil lokke til seg utanlandske produsentar, som dei som er nemnde ovanfor, til å lage norskspråklege produkt, til dømes generell diktering og maskinomsetjing. Når ein får på plass språkbanken, kan internasjonale selskap ta i bruk ressursane til å forbetre eller lage nye talegjenkjenningsprodukt for den norske marknaden. Dette kan gje oppdrag til norsk industri med omsyn til systemintegrasjon til dømes av taleteknologi i nye tenester og produkt.

Det er viktig å merke seg at EU-land som Frankrike, Italia, Nederland og Belgia no opnar for ein sterk offentleg innsats for å etablere språkteknologisk infrastruktur som mellom anna kan motstå presset frå engelsk språk. Tyskland vurderer liknande initiativ. Dette skjer altså i land *som frå før har ein mykje sterkare språkteknologisk infrastruktur enn Noreg*, og dette er store språksamfunn samanlikna med det norske. Heller ikkje desse landa har ein språkteknologisk industri som er sterk nok til å kunne bere kostnadene, men kulturpolitiske og næringsmessige prioriteringar gjer at desse landa nyttar store offentlege ressursar til å byggje opp nasjonale språkdatabasar. Eit lite språksamfunn som det norske kan ikkje rekne med å ha nokon sterk språkindustri med inntening som over tid kan finansiere kostnadene ved datainnsamlinga. Det

er nasjonen Noreg som må syte for å investere i nødvendig infrastruktur som gjer det mogleg å skape språkteknologiske produkt for norsk med same kvalitet som for andre språk.

Etter innsamlinga må ressursane haldast ved like, driftast og utviklast vidare. Brukarane, dvs. industri og forsking, må betale for bruken av innhaldet i samlinga, og noko av kostnadene med drifta bør etter kvart kunne dekkjast med slike vederlag. Prisen kan ikkje setjast særleg høg, då forsvinn heile poenget med at datasamlinga skal gjere det attraktivt å lage språkteknologiske produkt for norsk.

Prosjektgruppa ser to alternativ for finansiering av språkbasen:

a) full offentleg finansiering
b) det offentlege finansierer hovuddelen, aktuelle leverandører kan leggje inn data mot vederlag i form av andre data eller kontant oppgjer, og brukarer betaler ei mindre avgift for å bruke materiale frå samlinga

Ein bør opne for at industrien kan vere med og bestemme prioriteringane dersom han legg til midlar til innsamlinga. Dersom næringslivet skyt inn midlar, kunne ein gje høve til å leggje inn tidsklausular for konkurrerande selskap. Problemet med slike spleiselag er at eigarskapstilhøva blir kompliserte. Komplisert eigarskap og tidsklausular kan gjere det mogleg for einskilde aktørar å redusere eller blokkere konkurransen på marknaden, og det er ikkje ønskjeleg.

Det må vere eit absolutt krav at datasamlinga blir kvalitetskontrollert (validert) av ein nøytral instans og gjort tilgjengeleg for forsking og industri.

Samanlikna med dei prosjekta vi kjenner frå andre land der ein er i gang med innsamling, trengst det om lag 100 millionar kr for å få ei ressurssamling med eit innhald og ein kvalitet som gjer at ho kan brukast av moderne språkteknologisk industri og forsking. Midlane skal nyttast til å kjøpe fri eksisterande materiale med god nok kvalitet, til dømes frå Nordisk Språkteknologi, identifisere materiale eller datasamlingar som kan vidareformidlast, og finansiere nyinnsamling av data. Frikjøp og vidareformidling tek minst tid og bør prioriterast saman med nyinnsamling av spontan tale. Behovet for internasjonal kontakt, til dømes gjennom deltaking i EU-nettverket for språkressurssamlingar (*Enabler)*, må understrekast og prioriterast. Deltaking her vil gje innsikt i kva som skjer elles i Europa, og gjere det mogleg å utnytte internasjonal erfaring gjennom å følgje internasjonale standardar og den beste praksisen. Det betyr òg at ein kan få tilgang til å bruke og/eller leggje til rette programvare-verktøy som lettar innsamling og vidareforedling av språkdata. Resultatet kan bli lågare kostnader og betre kvalitet.

KAPITTEL 7 PLAN FOR GJENNOMFØRING

Planen for gjennomføring er ei skisse. Dei konkrete forslaga er ikkje diskutert i tilstrekkeleg grad med involverte institusjonar og aktuelle bransjeorganisasjonar. Dette har ikkje gruppa hatt tid til og forslaga må sjåast på som tentative og som skisser til moglege modellar.

7.1 Tidshorisont

Innsamling av ressursane bør kome i gang så raskt det lèt seg gjere. Ei spreiing over fem år må ein likevel rekne med fordi mykje av materialet må samlast inn frå grunnen, og ressursinnsamling er tidkrevjande. Når det gjeld finansieringa, er det også ein fordel at kostnadene blir spreidde over fleire år. Ein tidshorisont på dette nivået kongruerer godt med tilsvarande prosjekt i andre land.

7.2 Etableringskostnader og brukskostnader

Når ressurssamlinga skal etablerast, må ressursar kjøpast fri eller samlast inn på nytt for så å bli stilte til rådvelde for brukargruppene, dvs. industri og forskingsmiljø. Dei som nyttiggjer seg data, skal betale for ressursane, men ikkje på langt nær så mykje som det kostar å etablere ressurssamlinga (i så fall forsvinn heile poenget med å etablere ho). Til kommersiell bruk kan ein kunde betale til dømes 5 % av berekna kostnad for materialet i språkbanken, medan forskingsinstitusjonar bør sleppe med om lag halvparten av dette.

Frikjøp kan skje på fleire måtar:

1. Leverandøren får betalt ved innlevering til språkbanken
2. Leverandøren får tilbake verdiar til språkbankpris berekna etter inngangsverdien på dataressursane, dvs at ein byter til seg meir data enn ein legg inn
3. Leverandøren får tilbakebetaling etter kvart som data blir tekne i bruk
4. Ein betalingsmåte som kombinerer 1 og 2
5. Ein betalingsmåte som kombinerer 1 og 3.

Alternativ 2, 3, 4 og 5 kan vere ein måte å realisere eit spleiselag på, jamfør tabellen over kostnadsfordeling. Med ein slik prisstruktur vil dei som legg inn data etter alternativ 2, kunne få tilbake data til ein verdi som er mange gonger høgare enn det dei legg inn, noko som vil gjere det attraktivt å skyte inn ressursar.

Opphavsrettar og frikjøp av materiale må språkbankens styre sjå nærare på i kvart einskild sak. Generelle problemstillingar er utgreidde i den juridiske rapporten, og vi viser elles til punkt 7.13 i dette kapitlet.

7.3 Eksisterande materiale

Det er henta inn informasjon om aktørar som kan ha ressursar som kan takast inn i ressurssamlinga. Ei oversikt over aktørar og relevante ressursar er vist i vedlegga. Venteleg er det meir aktuelt materiale som kan inngå i samlinga, men oversikta etterlet eit inntrykk av at det eksisterer nokså mykje materiale som kan integrerast i ressurssamlinga.

Ein bør merkje seg at det finst til dels store mengder relevante ressursar som allereie er samla inn. Nokon av dei kan gå inn i databasen utan vidare. Men for mykje av materialet står det att kontroll både av kvalitet og i kva grad materialet i det heile kan nyttast av opphavsmessige

årsaker. Eit slikt arbeid er alt for omfattande og tidkrevjande til at det kan gjerast greie for innan dei gjevne tidsrammene, men nokre tentative tilrådingar er gitt i høgre kolonne i tabellane i vedlegga.

Verdifastsetjing av eksisterande materiale må baserast på estimat og kjende kostnader for nyinnsamling med dagens verktøy. Dersom det viser seg at nyinnsamling ikkje prismessig skil seg vesentleg frå priskrava, bør materialet samlast inn på nytt dersom det ikkje tek for lang tid.

Oppdragsgjevar må sjølv vurdere i kva grad ressursar som er samla inn over statlege budsjett utan vidare skal gjerast tilgjengelege for ressurssamlinga, gitt at opphavsrettslege vilkår er oppfylte. Dette gjelder særlig materiale som er samla inn ved universiteta. Meir om dette lenger ute i kapitlet.

7.3.1 Taledata

Det går fram av oversikta over eksisterande materiale (vedlegg 1, tabell x) at nokså store mengder manuskriptopplesen tale allereie er tilgjengeleg frå Nordisk språkteknologi. Det finst også omfattande mengder telefonopptak frå NST og Telenor. Spontan tale er nesten fråverande i dei eksisterande samlingane. Blir ein einige om refusjonsordningar, tilseier dette at ein kan frikjøpe manuskriptopplest materiale og realisere mykje av taledelen i språkbanken raskt. Innsamling av spontan tale må skje over fleire år.

7.3.2 Tekstdata

Ein del tekstdata ser ut til å vere tilgjengelege, men mykje av dette materialet er avgrensa til forskingsbruk. Nye data ein samlar inn på dette området, bør i størst mogleg grad vere kommersielt tilgjengelege, og tekstsamlinga må bli betre balansert med omsyn til teksttypar. Aviser og sakprosa er svært høgt representert.

7.3.3 Leksikalske data

Mengda av eksisterande leksikalske data er jamt over akseptabel, med eit par unntak. Men ein må rekne med ein relativt stor innsats når det gjeld opprensking og systematisering av samlingane, bl.a. ved at grammatisk merking og uttalekonvensjonar blir standardisert. Uansett er slikt standardiseringsarbeid mindre omfattande enn innsamling av nytt materiale med tilhøyrande merking og generering av bøyingsparadigme.

Når det gjeld uttalebeskrivingar må ein også standardisere og gå gjennom materialet, både med omsyn til annoteringskonvensjonar (SAMPA, XSAMPA, osv), markering av stavingsgrenser, trykk, tonelag og dialektopphav.

7.3.5 Tilråding

Ein bør i størst mogleg grad nytte eksisterande ressursar om dei har tilstrekkeleg kvalitet. Dei opphavsrettslege tilhøva må vere avklarte. Alle data som er aktuelle, må kontrollerast av nøytrale ekspertar. Kompensasjonsordningar må diskuterast med kvar enkelt leverandør.

7.4 Ressursar finansierte over staten sine ordinære løyvingar

Her finn ein dei viktigaste kjeldene ved universiteta i Bergen (UiB) og Oslo (UiO), og noko ved NTNU. I Bergen har ein tekst- og ordlistemateriale ved HIT-senteret (tidlegare NAVFs Edb-senter for humanistisk forsking), jamfør vedlegga om eksisterande materiale. Universitetet i Oslo har tekst og ordlistemateriale av verdi for språkbanken. Følgjande

ressursar er identifiserte og verdien er fastsett etter prinsippa i tidlegare kapittel (ein føreset at data har tilstrekkeleg kvalitet):

Tabell 7.1: Aktuelle ressursar eigd av universiteta

Inst.	*Ressurstype*	*Total verdi pr institusjon*
UiB:	Tekstar, 0,5 mill	0,5
NTNU:	Tale, leksikalske data: 0,8 mill	0,8
UiO:	Tekstar: 1,9 mill, leksikalske data: 6,6 mill, taledata: 1 mill	9,5 mill (reknar ikkje inn NorKompLeks[1] frå NTNU pga overlapping, går ut frå at det er like mykje nynorsk som bokmål)

Læringssenteret (UFD) har noko dei kallar Lydbokbanken med innlesen digital tale av god kvalitet. Denne ressursen blei prosjektgruppa merksam på så seint i arbeidet at det ikkje har vore mogleg å vurdere materialet nærare, men det bør gjerast. Læringssenteret formidlar òg DAISY-plater. (DAISY = Digital Accessible Information System). Dette er CD-ROM-plater med inntil 50 timar lyd. Dei blir vanlegvis brukte til lærebøker for syns- og hørselshemma, og dei kan spelast av på vanlege PC-ar med tilpassa utstyr (krev ekstrautstyr). DAISY er i ferd med å bli ein internasjonal standard, og neste versjon vil liggje tett opp til eit digitalt, multifunksjonelt format som det er teknisk mogleg å tilby over nettet.

Språkbanken bør forhandle med dei aktuelle leverandørane, særleg Universitetet i Oslo, om vilkår for innleming av desse språkressursane i basen.

7.5 Ressursar finansierte av andre verksemder i staten sitt eige

Telenor er her den mest aktuelle leverandøren. Ressursane Telenor har, er taledata og transkriberte ordlister (overslaga er usikre).

Tabell 7.2: Aktuelle ressursar hos Telenor

Ordlister	0,2 mill. (overslag på 50 000 norske ord frå Onomastica[2])
Taledata	2 mill. (truleg meir som ikkje kan kvantifiserast med tilgjengeleg informasjon)

Det er urealistisk å vente at Telenor vil overdra data sine utan vederlag, men gruppa vil foreslå at staten prøver å finne ei løysing som tener språkbanken sine kortsiktige behov, til dømes ved at Telenor får tilbake andre ressursar som dei har behov for, eller at ressursane blir betalte over tid etter kvart som språkbanken kjem i ordinær drift.

7.6 Ressursar finansierte (heilt/delvis) av Noregs forskingsråd

Dette gjeld i første rekkje ressursar som no må samlast inn via forskingsprogrammet KUNSTI, fordi nødvendige språkressursar ikkje er tilgjengelege. Sidan elementa i dette programmet i skrivande stund ikkje er klarlagde, kan ein ikkje rekne med data herfrå, men i kontraktane som skal skrivast, bør det stå at innsamla data skal overlatast til språkbanken. I tillegg kjem språkteknologiske verktøy som er utvikla med finansiering frå Noregs forskingsråd, til dømes program for automatisk ordklassemerking ved Universitetet i Oslo og HIT-senteret. Slike verktøy kan språkbanken nytte vederlagsfritt, men arbeidskostnadene med merkingsarbeidet må dekkjast.

[1] Norsk ordliste med informasjon om bøying, valens og uttale.
[2] EU-prosjekt der ein transkriberte lydskrift av forventa uttale av fornamn, etternamn, stadnamn osv. i 11 europeiske land, i alt 8,5 millionar namn.

7.7 Ressursar via verkemiddelapparatet

Data til Nordisk Språkteknologi er dei mest relevante i denne kategorien. Nordisk Språkteknologi har fått og får mykje støtte via offentlege finansieringskjelder, men har ikkje fått tilskot til innsamling av sjølve språkdata. Prosjektgruppa ser desse data som verdfulle å ha i ressurssamlinga.

Ein kan ikkje vente at Nordisk Språkteknologi kan leggje ressursane sine inn i ressurssamlinga om ikkje selskapet får noko igjen for det. Staten bør likevel vurdere om det er rimelig å be om rabatt for materialet frå Nordisk Språkteknologi.

Tabell 7.3: Aktuelle tilleggsressursar

Leksikalske data utover materialet frå Norsk ordbank ved UiO	Ca. 1 mill.
Taledata	12,1 mill.

7.8 Andre data som kan inngå i språkbanken

Tekstmateriale frå BerlitzGlobalNet og Oracle kan vere aktuelt for ressurssamlinga, jamfør vedlegg 1. Desse aktørane har signalisert at dei som kompensasjon ønskjer andre språkressursar. Verdien av materialet er berekna til 3,6 millionar kr. Forlaga har også ordbøker og tekstmateriale som kan vere interessant for språkbanken. Prosjektgruppa har vore i kontakt med Kunnskapsforlaget og Det Norske Samlaget som begge stiller seg positive til å hjelpe til under føresetnad av at nødvendige avtalar og kontraktar kjem på plass.

7.9 Kostnadsdeling under innsamling

Universiteta i Bergen og Oslo har signalisert vilje til å hjelpe til med personalressursar i innsamlingsarbeidet. Dette er særleg viktig for den faglege delen som er knytt til nyinnsamling. Kor omfattande bidraga frå universiteta kan bli, vil avhenge av kva for innsamlingsprosjekt ein legg til universiteta, og om og i kva grad den aktuelle kompetansen er tilgjengeleg.

For Universitetet i Oslo er det mest aktuelt med arbeid knytt til leksikalske ressursar. Universitetet i Oslo har ikkje gjeve signal som er tydelege nok til at ein kan talfeste innsatsen, men bør kunne rekne med minst éin stillingsressurs per år i fire år. Føresetnaden er at institusjonen får oppgåver som passar til kompetanseprofilen. Verdien for språkbanken er omlag 2,5 - 3 millionar kr, meir om større ressursar blir stilte til disposisjon.

Universitetet i Bergen har signalisert interesse for arbeid med tekstsamlingar, og har konkretisert dette til om lag 2 stillingar knytt til dette arbeidet. Eit samarbeid med Universitetet i Oslo verkar naturleg i denne samanhengen. Vi kan kalkulere med rundt to stillingar på forskarnivå eller leiarnivå i heile innsamlingsperioden på fem år. For Universitetet i Bergen gjeld også at prosjekta må passe til kompetansen. Verdien av desse stillingane er mellom 6,5 og 8 millionar kr.

Ein annan type kostnadsdeling ved nyinnsamling av materiale gjeld bidrag som følgje av prioriterte OFU- eller IFU-prosjekt via SND.

Formålet med OFU-ordninga er å styrkje næringslivets konkurranseevne både nasjonalt og internasjonalt gjennom samarbeid med ein krevjande offentleg kunde. Midlane skal stimulere

til å betre kvaliteten og/eller redusere kostnadene på offentlege tenester gjennom tilgang til ny teknologi eller nye løysingar.

Formålet med IFU-ordninga er å stimulere til FoU-samarbeid mellom kundebedrifter og leverandørbedrifter om utvikling av nye prosessar, metodar eller tenester som ei eller fleire bedrifter kan nyttiggjere seg. Vidare skal ordninga medverke til å utvikle konkurransedyktige produkt med eksportpotensial, gjerne i samarbeid md ei utanlandsk kundebedrift.

Eit døme på kor nødvendig det er å ha språkdata tilgjengelege, er det som no skjer i eit prosjekt med såkalla medisinsk diktering. Dette er eit OFU-prosjekt mellom Nordisk Språkteknologi, St. Olavs Hospital og SND, og hadde ikkje vore mogleg utan tilgang til dei språkressursane som Nordisk Språkteknologi alt hadde samla inn. Resultatet frå dette prosjektet kan danne grunnlag for andre dikteringsprosjekt, og nye bruksområde og -måtar. Dette, saman med tilgang til fleire og andre språkressursar frå ein språkbank, kan nyttast i andre prosjekt med offentlege etatar (OFU) eller bedrifter (IFU), dersom relevante data blir stilte til rådvelde for språkbanken.

I OFU/IFU-prosjekt må partane bli einige om at dei data som blir utvikla, kan stillast til disposisjon for språkbanken prissett til den summen tilskotet frå SND utgjer, eventuelt kan ein avtale dette særskilt i kontrakten.

Det kan vere vanskeleg å talfeste verdien av dette, men ut frå dei kontraktane som er inngått, kan vi rekne med verdiar tilsvarande 2 millionar per år i den perioden SND eventuelt prioriterar språkteknologi i støtteordningane sine.

7.10 Finansieringsmodell

Nedanfor står ei skisse til ein modell for finansiering av ressurssamlinga. Ein legg til grunn ei ordning med å kjøpe fri eksisterande. Etter at språkressursane er lagt inn i språkbanken, kan ressurane kjøpast for opptil 10 % av det det har kosta å leggje dei inn. 10 % av innkjøpskostnadene er i høgaste laget, og vil prise dei norske ressursane til monaleg over ELRA-ressursar. Prisnivået for materiale frå språkbanken må vere ein god del under 10 % av innkjøpskostnadene. Spesielt gjeld dette for bruksområde som krev store ressursmengder, til dømes dikteringssystem. Som peikt på andre stader i rapporten: Dersom prisnivået blir for høgt, forsvinn heile poenget med å etablere ressurssamlinga. Forskingsinstitusjonar bør, i tråd med etablert praksis i ELRA/ELDA, betale mindre enn kommersielle aktørar.

I kolonnen *Totalkostnad* er den estimerte verdien av data ført opp, og dei er splitta i dei tre hovudkategoriane taledata, tekstdata og leksikalske data. Gruppa presiserer at *validering* av data er inkludert i totalkostnadene, jamfør kapitlet om innhaldet i samlinga. Kolonnen *Eksisterande data* refererer til data som er tilgjengelege og sannsynlegvis kan inngå i basen, jamfør omtalen av dei tilgjengelege ressursane (kapittel 7.3). Desse data kan bytast mot andre data etter kvart som dei kjem inn i basen. Men dersom ein leverandør legg større ressursar inn i samlinga enn verdien av ressursane som vedkomande har interesse av å hente ut, kan ein ikkje rekne med at alle data blir stilte til rådvelde utan vederlag. Kolonnen *Avgifter* viser til avgifter kundane betaler til basen for bruk av ressursane. Det er realistisk å rekne med til dels stor overlapping mellom dei som leverer data til språkbanken og dei som vil kjøpe data frå han, og det er grunn til å tru at dei fleste leverandørane vil ha kontant oppgjer. Potensialet for innsamling av eksisterande data er derfor redusert med 50 %. Kolonnen *Kostnadsdeling - nyinnsamling* viser til stipulert eigeninnsats frå universiteta i samband med innsamling av nye ressursar.

Tabellen nedanfor er eit optimistisk anslag:

Tabell 7.4: Kostnadsoverslag I

Type	Totalkostnad	Eksisterande data	Avgifter	Kostnadsdeling – nyinnsamling	Netto-kostnad
Taledata	46 mill.	7 mill.	2 mill.		
Tekstdata	30 mill.	3 mill.	2 mill.	8 mill.	
Leksikalske data	16 mill.	3 mill.	1 mill.	3 mill.	
Administrasjon	7 mill.				
Sum	99 mill.	13 mill.	5 mill.	11 kr mill.	70 mill.

Vi kalkulerer med at 5 % av verdien på materialet i samlinga kan kome frå avgifter. Denne verdien vil bli realisert over tid, kan hende over heile innsamlingsperioden. Staten og leverandørar må kanskje vente på at inntekter blir realiserte. Denne situasjonen er velkjend frå andre ressurssamlingar, jamfør ELRA. Vi tek ikkje stilling til om inntektene blir implementerte som medlemsavgifter åleine (jamfør LDC), eller som ein kombinasjon av medlemskap og kjøp av aktuelle ressursar (jamfør ELRA).

Det er optimistisk å rekne med at ein kan realisere eksisterande data til ein verdi av 13 millionar kr etter innbytemodellen, og derfor er dette ein usikker parameter i modellen. Eit nivå som kanskje er meir realistisk, er halvdelen av dette.

Prosjektgruppa meiner modellen kan fungere bra sjølv om ressursane får ein prislapp som i verste fall kan leie til at somme produkt ikkje blir utvikla for norsk språk. Men om alt innhaldet i basen er gratis, vil aktørar som har relevant materiale, neppe ta bryet med å stille det til rådvelde for basen. At verdien av innlevert materiale er kopla til kva ein får ut, bør vere eit insitament til dei som gjev data til språkbanken. Dersom ein leverandør får verdsett sin leveranse, kan det kome til fråtrekk når han kjøper ressursar frå banken. Om ei verksemd legg inn ressursar tilsvarande ein verdi på 2 millionar kr, vil ho få tilbake verdiar som minst svarar til 20 millionar kr.

Høg pris på ressursane vil hemme industriell satsing og utvikling av språkteknologi for norsk. Verdien av eigendelane blir realiserte over tid, og det vil i praksis fungere som ein finansiell basis for drift og vedlikehald av språkbankorganisasjonen. Balansen ein har funne fram til her, er kanskje i høgste laget. Eit meir realistisk alternativ som tek opp i seg reservasjonane ovanfor, er dette:

Tabell 7.5: Kostnadsoverslag II

Type	Totalkostnad	Eksisterande data	Avgifter	Kostnadsdeling – nyinnsamling	Netto-kostnad
Taledata	46 mill.	3 mill.	1 mill.		
Tekstdata	30 mill.	1 mill.	1 mill.	8 mill.	
Leksikalske data	16 mill.	1 mill.	1 mill.	3 mill.	
Administrasjon	7 mill.				
Sum	99 mill.	5 mill.	3 mill.	11 kr mill.	80 mill.

7.11 Budsjett

I det svært tentative budsjettet frå Delrapport 1 såg ein for seg at første år legg beslag på 10 % av totalkostnadene, 30 % i det andre året, og 20 % for kvart av dei følgjande tre åra. Motivasjonen er at materiale ein utan vidare kan innleme, bør kvalitetssikrast og inkluderast i løpet av dei første to åra. Alt aktuelt materiale må vurderast grundig. Denne vurderinga bør ein gjere i første driftsåret med innleming i påfølgjande år.

I all innsamling er det rekna med midlar til forskarar og assistentar. Oppgåvene for forskarane er å spesifisere, leie og kvalitetssikre innsamlingane. Andre forskarar skal ta seg av kvalitetssikringa, men begge typane forskingsoppgåver er inkluderte i kalkylane for innsamlingsarbeidet. Denne modellen fordrar at forskingsinstitusjonar, institutt, kompetente bedrifter eller utanlandske institusjonar (til dømes nederlandske SPEX) deler arbeidet med innsamling og kvalitetssikring mellom seg (utanlandske institutt kan berre ta seg av kvalitetssikringa). Forskingsarbeidet er i snitt estimert til 25 % av innsamlingskostnadene.

Når det gjeld frikjøp av eksisterande materiale, blir det taksert ut frå kostnadene med nyinnsamling. Kvalitetssikring av data er innbakt i kostnadsoverslaget.

7.12 Administrasjon

7.12.1 Administrasjon i etableringsfasen

Det administrative arbeidet er mest ressurskrevjande dei første par åra. Då skal eksisterande materiale evaluerast for mogleg innkjøp (teknisk, innhaldsmessig og juridisk), detaljerte spesifikasjonar for nyinnsamlingar skal utarbeidast, og innsamling av nytt materiale skal setjast ut på anbod.

7.12.2 Administrasjon etter innsamlinga

Etter kvart som språkbanken får inn ressursar som er kvalitetssikra, kan materialet distribuerast av til dømes ELRA på vegner av språkbankorganisasjonen. Dette utelukkar ikkje at andre aktørar enn ELRA kan stå for distribusjonen nasjonalt. Prosjektgruppa ser det som ønskjeleg at språkbankorganisasjonen sjølv eller det foreslåtte driftsselskapet tek hand om distribusjonen i Noreg. I forlaget til statsbudsjett for 2003 frå NHD, står det i resultatrapporten for 2001 følgjande: ”I 2001 stiftet SPNE sammen med Voss kommune selskapet EDDA Språkressurser as. Dette selskapet skal utvikles i retning av en nordisk språkbank, og vil kunne tilføre oppstartsbedriftene i inkubatoren [SPNE] nyttige tjenester og kompetanse innen språkteknologi.” (St.prop.nr. 1 2003-2004, kapittel 927 Språkteknologisenter, s. 126). Ein bør ta dette miljøet med i vurderinga med omsyn til kven som kan gjere kva av eksisterande kompetente fagmiljø.

Organisasjonsmodellen opererer med ei driftseining som fungerer på vegner av stiftinga sitt styre. Økonomien for einings bør vere sjølvfinansierande ved at det blir kravd inn ei avgift for alt materiale som blir levert ut.

Permanente driftskostnader bør delast med eksisterande einingar ved universiteta. Universitetet i Oslo har einings Norsk ordbank under etablering, og banken kan til dømes få spesialansvar for drift og vedlikehald av dei leksikalske ressursane etter at innsamlingsperioden er over. HIT-senteret ved Universitetet i Bergen har lang erfaring med distribusjon av tekstsamlingar, og denne kompetansen kan nyttast ved at HIT-senteret tek ansvar for

vedlikehald av tekstressursane, gjerne i samarbeid med Tekstlaboratoriet ved Universitetet i Oslo. Begge institusjonar har signalisert interesse for ei slik løysing.

Når det gjeld taledata, kan ein sjå for seg ei løysing der anten Universitetet i Oslo eller i Bergen tek hand om forvaltinga av dei.

7.13 Leveringsplikt

Prosjektgruppa reknar med at eit fleirtal av verka som skal inngå i språkbanken, er opphavsrettsleg verna åndsverk. Bruk av verket til eksemplarframstilling eller anna tilgjengeleggjering av verket er avhengig av samtykke frå opphavleg eller avleidd rettshavar. Elektronisk lagring av verket er avhengig av samtykke, særleg etter implementering av opphavsrettsdirektivet i norsk rett. Opphavsrettar eller bruksrettar kan overtakast gjennom avtale med forvaltingsorganisasjonen LINO, som er oppretta for slike formål.

I tillegg kjem all offentleg informasjon i form av utgreiingar, meldingar, lovverk osv. Føresetnaden er naturleg nok at materiale som blir inkorporert i språkbanken, ikkje skal kunne republiserast av språkbankkundar, heller ikkje som tekstarkiv for informasjonsgjenfinning. Med ei slik leveringsplikt vil ein sleppe frikjøp for mange av språkbanken sine tekstsamlingar, og kostnadane kan såleis haldast nede. Leveringsplikta kan bli administrert av Nasjonalbiblioteket i samråd med språkbanken sin driftsorganisasjon.

Dette betyr ikkje at alt skriftleg materiale omfatta av ei slik ordning skal inngå i språkbanken, men leiinga i språkbanken får med denne løysinga sjansen til å balansere korpusa med omsyn til målform, sjanger, periode, forfattarar, osv., noko som elles er eit stort problem for nesten alle språkteknologisk motiverte tekstsamlingar.

I innsamlingsscenaria er det føresett at tekstmaterialet som skal samlast inn, blir underlagt ei ordning med leveringsplikt eller andre løysingar som minimaliserer eventuelle leveringsgebyr.

Eit innspel frå Faglitterær forfatterforening viser at ei ordning med innlevering av tekstmateriale kan la seg gjennomføre utan altfor store problem:

"Rettighetsklarering knyttet til lagring og senere bruk av åndsverket innebærer i utgangspunktet en to-trinns prosess, hvor det må inngås selvstendige avtaler for hhv. lagringen og bruken av verket (evt. kollektiv avtale gjennom LINO). Administrasjonen knyttet til lagringsrettighetene vil imidlertid kunne lettes ved at det etableres en avleveringsplikt for en nærmere avgrenset gruppe åndsverk. En slik ordning vil være naturlig å knytte opp mot dagens pliktavleveringslov av 9. juni 1989 nr. 32. Nærmere angivelse av hvem som omfattes av avleveringsplikten, den praktiske gjennomføring etc. vil kunne reguleres i forskrift, jf. l. § 5 4. ledd.

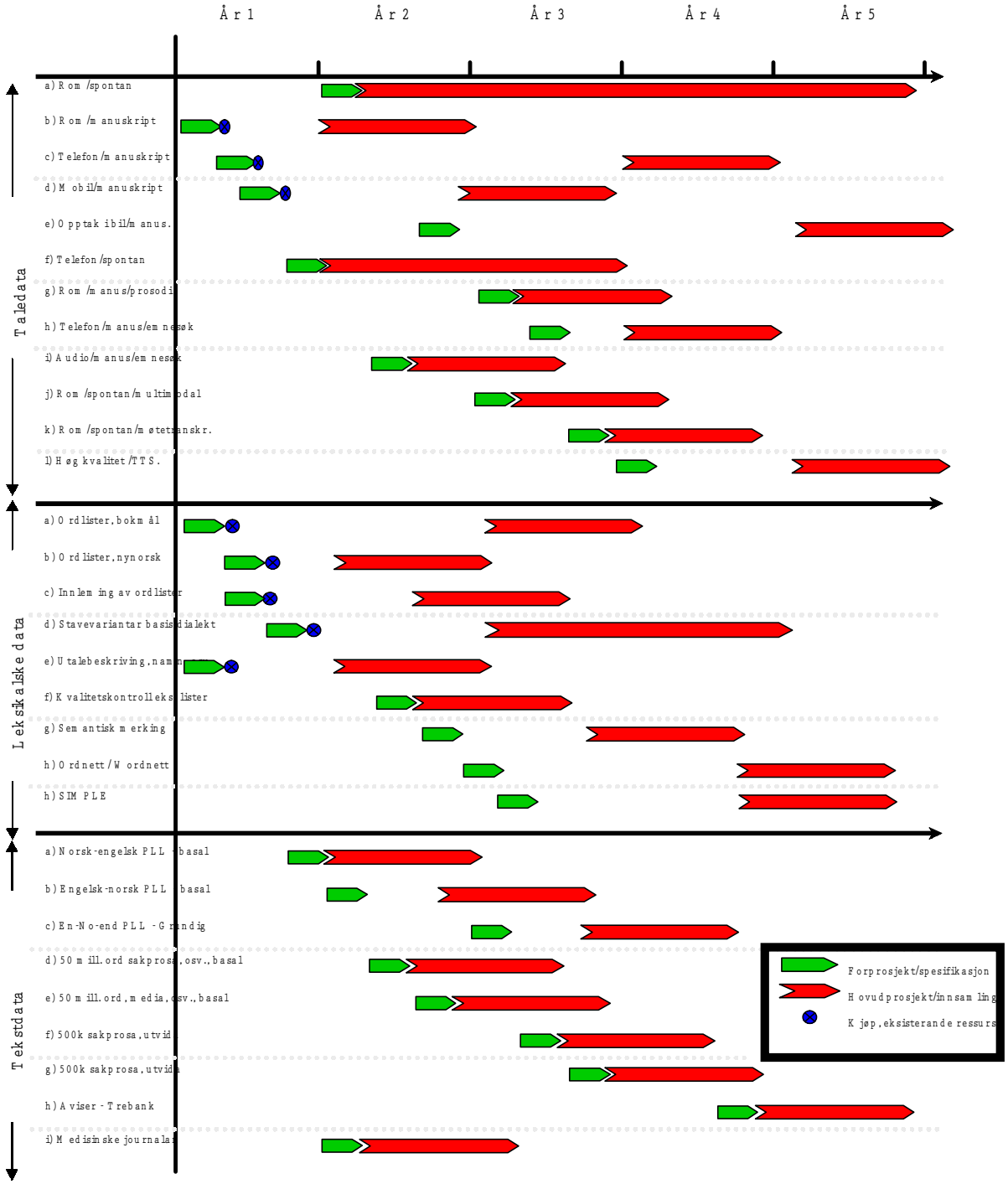
Hensynet bak lov om pliktavlevering er et kulturelt begrunnet behov for å bevare og dokumentere de verk som skapes i samfunnet. Hensynet til opphavsmannens enerett til lagring (eksemplarfremstilling) av verket skaper i denne sammenheng neppe noe motsetningsforhold. Hvis de pliktavleverte verkene senere skal gjøres tilgjengelige for allmennheten, reguleres dette av åndsverkloven og eventuelle avtaler, jf hva som er sagt om administrasjon i punkt 6.1.

En mangel på tekniske standarder har tidligere vært et problem ved administrasjon av avleveringsplikten. Det vil derfor være hensiktsmessig at det enkelte åndsverk avleveres på samme digitale plattform. I den sammenheng er det viktig at plattformen er i samme digitale format som språkbanken anvender. De eventuelle merkostnader en slik standard måtte innebære for den avleveringspliktige, vil kunne dekkes helt eller delvis av mottakerinstitusjonen (språkbanken), jf l. § 5 2. ledd."

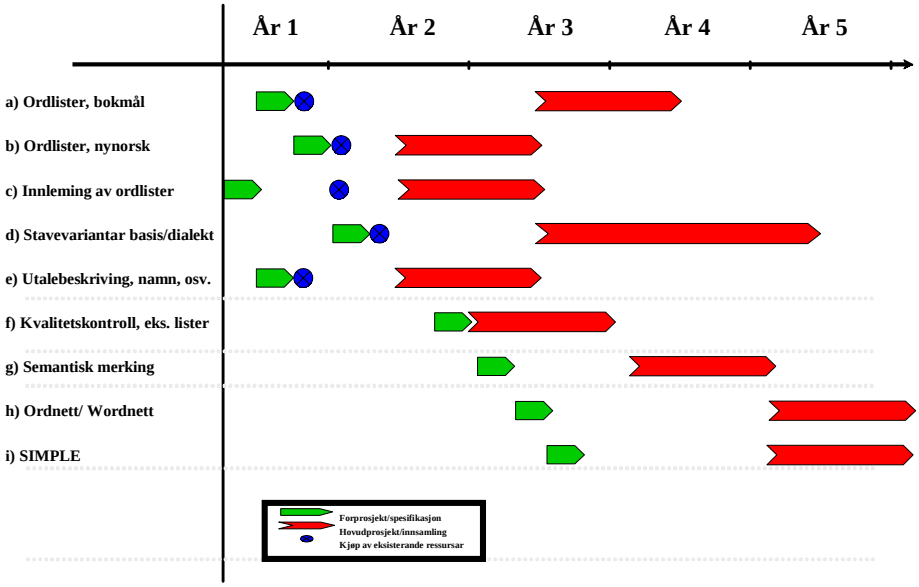
Det må bli opp til språkbanken sitt styre og administrasjon å utarbeide detaljane for ei ordning med pliktavlevering til språkbanken.

7.14 Tidsplan for innsamlinga

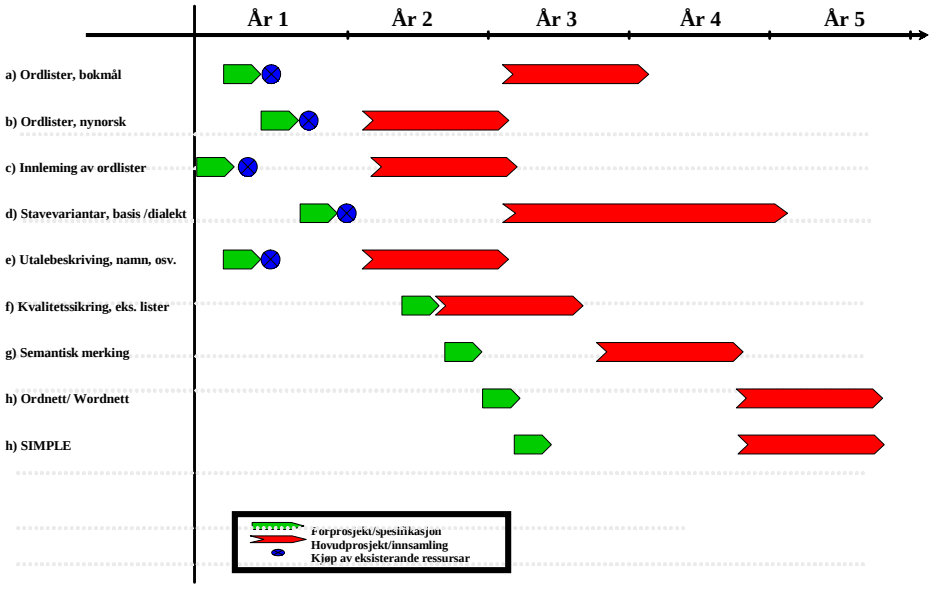
På dei følgjande sidene er det skissert ein tidsplan for innsamlinga. For alle modulane er det sett av ressursar til forprosjekt, gjennomføring og kvalitetssikring av data (kvalitetssikringa er inkludert i hovudprosjekta). Planen er tentativ og må justerast i takt med tildelingane til samlinga, og eventuelle andre forhold som kan påverke prioriteringane. Slike justeringar må språkbanken sitt styre ta seg av.


År 1
År 2
År 3
År 4
År 5
Taledata
a) Rom /spontan
b) Rom /manuskript
c) Telefon/manuskript
d) Mobil/manuskript
e) Opptak i bil/manus.
f) Telefon/spontan
g) Rom /manus/prosodi
h) Telefon/manus/emnesøk
i) Audio/manus/emnesøk
j) Rom /spontan/multimodal
k) Rom /spontan/møtetranskr.
l) Høg kvalitet/TTS.
Leksikalske data
a) Ordlister, bokmål
b) Ordlister, nynorsk
c) Innlemming av ordlister
d) Stavevariantar basis/dialekt
e) Utalebeskriving, namn
f) Kvalitetskontrollekslister
g) Semantisk merking
h) Ordnett/Wordnett
h) SIMPLE
Tekstdata
a) Norsk-engelsk PLL - basal
b) Engelsk-norsk PLL - basal
c) En-No-end PLL - Grundig
d) 50 mill. ord sakprosa, osv., basal
e) 50 mill. ord, media, osv., basal
f) 500k sakprosa, utvida
g) 500k sakprosa, utvida
h) Aviser - Trebank
i) Medisinske journalar
Forprosjekt/spesifikasjon
Hovudprosjekt/innsamling
Kjøp, eksisterande ressurs

Leksikalske data



Tekstdata



FORKORTINGAR BRUKTE I RAPPORTEN

BNC	British National Corpus
DAISY	Digital Accessible Information System
EAGLES	Expert Advisory Group for Language Engineering Standards
ECI	European Corpus Initiative
ELRA/ELDA	European Language Resource Assosiation / European Language Distribution Agency
Enabler	Europeisk nettverk for språkteknologi
FoU	Forsking og utvikling
HIT-senteret	Senter for humanistisk informasjonsteknologi
IKT	Informasjons- og kommunikasjonsteknologi
IKT-Norge	Medlemsorganisasjon for IKT-bedrifter i Noreg
IT	Informasjonsteknologi
KKD	Kultur- og kyrkjedepartementet
KUNSTI	Kunnskapsutvikling for norsk språkteknologi
LDC	Linguistic Data Consortium (USA)
LINO	Organisasjon som forvaltar avtalar om opphavsrettar
LS	Læringssenteret, organ under Utdannings- og forskings-departementet (UFD)
NFR	Noregs forskingsråd
NHD	Nærings- og handelsdepartementet
NTNU	Noregs teknisk-naturvitskapelege universitet
NST	Nordisk Språkteknologi AS
OFU-/IFU-prosjekt	Stønadsordningar til industriutvikling (frå SND)
SAMPA/XSAMPA	System for merking av språkressursar
SIVA	Selskapet for industrivekst
SND	Statens nærings- og distriktsutviklingsfond
SOU	Svensk offentleg utgreiing
SPEX	Nederlandsk firma som kvalitetssikrar språkressursar
SPNE	S.AI.L. Port Northern Europe AS
TEI	Text Encoding Initiative
UFD	Utdannings- og forskingsdepartementet

SENTRALE DOKUMENT OG UNDERLAG FOR RAPPORTEN

Anbefalinger fra arbejdsgruppen IT på dansk, Ministeriet for Videnskab, Teknologi og Udvikling, Danmark 2001

BLARK: Definisjon av innhaldet i den nederlandske taleressursbasen, i D. Binnenpoorte o.a.: *A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch*, innlegg på LREC 2002, (Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spania, mai-juni 2002

eContent Rådsvedtak av 22. desember 2000 om vedtakelse av et flerårig fellesskapsprogram for å stimulere til utvikling og bruk av europeisk digitalt innhold i verdensomspennende nett og for å fremme språklig mangfold i informasjonssamfunnet (301D0048, uoffisiell norsk oversettelse) – EU-program for 2001-2004

eEurope *An Information Society for All* – EU-kommisjonens handlingsplan, 1999 og senere (motsvares av eNorge)

eNorge, handlingsplanar, versjonane 1.0, 2.0, 3.0, NHD (den norske regjeringens motstykke til eEurope)

eNorge 2005, NHD 2002

Handlingsplan for norsk språk og IKT, revidert utgåve, Norsk språkråd 2001

INFO2000 *Rådsvedtak av 20. mai 1996 om vedtakelse av et flerårig fellesskapsprogram som skal stimulere utviklingen av en europeisk multimedie-innholdsindustri og fremme bruken av multimedie-innhold i det framvoksende informasjonssamfunnet* (396D0339) – EU-program for 1996-2000 (videreført gjennom eContent)

KUNSTI – Kunnskapsutvikling for norsk språkteknologi, programplan, NFR 2001

MLIS *Rådsvedtak av 21. november 1996 om vedtaking av eit fleirårig program for å fremje språkmangfaldet i Fellesskapet i informasjonssamfunnet* (396D0664) – EU-program for 1995-99 (videreført gjennom eContent)

Mål i mun, Förslag till handlingsprogram för svenska språket, SOU 2002:27, Stockholm 2002

Norge – en utkant i forkant. Næringsrettet IT-plan 1998-2001, NHD februar 1998

Norsk språkbank. Utredning om et nasjonalt korpus for språkteknologi, Svendsen o.a. 1999

Plan for styrking av norsk, Norsk språkråd 2001

Satsing på informasjonsteknologi for funksjonshemmede (IT-FUNK), 1998-2001, NFR 1998

"Si@!". Elektronisk samhandling i helse- og sosialsektoren, statlig tiltaksplan 2001-2003, SHD 2001

Språkteknologi i Norge – eksisterende og påkrevet forskning, rapport, NFR 2000

St.meld. nr. 9 (2001-2002) Målbruk i offentleg teneste

St.meld. nr. 13 (1997-98) *Målbruk i offentleg teneste*

St.prp. nr. 1 (2002-2003), oktober 2002, for KKD og NHD

Strategi for elektronisk innhald 2002 – 2004, NHD 2002

Strategi for eksport og internasjonalisering av IKT-næringen, NHD 2001

Strategi- og handlingsplan for IKT-forskningen i Forskningsrådet, NFR 2000

Strategisk plan for Norsk språkråd 2000-2003, Norsk språkråd 2000

Taleforbedring for funksjonshemmede, sluttrapport, SINTEF 1999

VEDLEGG 1: Materiale som bør vurderast i samband med språkbanken

Nedanfor står eit tabellarisk oversyn over materiale som kan tenkast å gå inn i språkbanken. I den første kolonnen er det informasjon om kva slags data det er snakk om, deretter kjem namn på leverandørinstitusjonen, kalkulert verdi på materialet (i den grad det har vore mogleg med grunnlag i den informasjonen prosjektgruppa har hatt), i kva grad materialet er tilgjengeleg for ressurssamlinga, kva kompensasjonsordningar som kanskje kan brukast (dei fleste har ikkje skrive noko om dette), målform, omfang på materialet og til slutt ein generell kommentar. Gruppa vil understreke at verdianslaga er svært usikre.