Sikring og tilgjengeliggjøring av nordiske språkressurser (SITS)

1 Sammendrag

Behovet for en norsk språkbank har lenge vært erkjent i norske språkteknologiske fagmiljøer innenfor akademia og næringsliv. En språkbank er nødvendig for å sikre norsk språk en fremtid i den digitale tidsalderen, til beste for nordisk språkforsking og for kommersiell utvikling av programvare for prosessering av tale eller tekst. Selskapet Nordisk språkteknologi Holding AS gikk konkurs i november 2003. Det har lenge vært uavklart hva som skulle skje med de omfattende språkressursene som selskapet bygget opp. Nå er dette imidlertid avklart, da en gruppe av fem norske aktører har gått sammen og kjøpt ressursene fri fra boet. Kjøpergruppen består av Språkrådet, Universitetet i Bergen, Universitet i Oslo, NTNU og IBM Norge. Gjennom dette kjøpet er grunnlaget lagt for en etablering av en norsk språkbank.

Prosjektet har som mål å sørge for at dette omfattende og varierte materialet blir forsvarlig sikret og tilgjengeliggjort for brukere. Dette arbeidet vil umiddelbart komme brukermiljøene til gode, da de er sikret tilgang til et svært omfattende materiale bestående av akustiske og leksikalske databaser og tekstkorpus som er verdifulle både som forskningsressurser og som grunnlag for utvikling av språkteknologiske produkter og løsninger.

Per i dag består materialet av et stort antall datafiler i ulike formater og ulike lagringsformer (server, CD-ROM, DLT-tape). Målet for prosjektet er å utføre arbeidsoppgaver som er umiddelbart nødvendig for å sikre ivaretaking av ressursene fra NSTs konkursbo, og å systematisere og strukturere dataene slik at de fremstår som en velorganisert og oversiktlig infrastruktur for forskning og utvikling. Dette innebærer følgende delmål:

- sikring av dataene, ansvarlig lagring, sikkerhetskopiering
- systematisering, fjerning av duplikater og irrelevant materiell, versjonskontroll
- strukturering av data, metainformasjon i henhold til internasjonal standard
- system for gjenfinning av data
- etablere avtaleverk for lisensiering til brukere
- nettsted med innholdstjenester, søkbar ressurskatalog, dataprøver
- informasjons- og formidlingsoppgaver, informasjonsmøter osv.

Oppgavene vil bli utført på en slik måte at prosjektet resulterer i et best mulig utgangspunkt for en fremtidig norsk språkbank. Dette arbeidet vil være av nasjonal betydning. Gjennom en rekke utredninger det slått fast at å etablere en språkbank er en nasjonal oppgave, senest i St.meld. nr. 17 (2006-2007), hvor frikjøpet av ressursene er eksplisitt nevnt som et første steg mot en slik etablering. Kostnadene med å etablere en nasjonal språkbank har blitt betydelig nedjustert etter dette oppkjøpet. Det er nå nødvendig å sikre at denne offentlige investeringen fører til økt verdiskaping ved at ressursene kommer brukermiljøene til gode. Gjennom prosjektet sikres norske brukermiljøer tilgang til en nasjonal forskningsinfrastruktur som vil være en betydningsfull komponent i en fremtidig norsk språkbank, når et vedtak om dette foreligger. Prosjektet har også en nordisk og internasjonal verdi, da materialet inneholder betydelige språkressurser også på dansk og svensk. Behovet for at et slikt tiltak anses som stort, og det har vært mange nylige henvendelser fra aktuelle brukere i inn- og utland.

2 Bakgrunn

En språkbank er en samling av språkressurser i form av store mengder tekst og tale som er elektronisk lagret og tilrettelagt for ulike typer bruk og gjenbruk, særlig tenkt brukt til

utvikling av språkteknologiske produkter. Språkressurser omfatter primærressurser som tekstdatabaser (korpus), taledatabaser og leksikalske databaser og sekundærressurser som språkbehandlingsverktøy/programvare som er utviklet på grunnlag av primærressurser, samt kunnskapssystemer som grammatikker, ordnett, terminologi, ontologier osv.

Norske fagmiljøer som driver med forskning og utvikling innen språkteknologi har behov for omfattende språkressurser, som representerer den språklige bredden som karakteriserer Norge som språksamfunn. Oppbygging av slike ressurser er kostbart, og det har gjennom en rekke utredninger vært slått fast at oppbygging av slike ressurser er en nasjonal oppgave. Det nasjonale ansvaret for er fremhevet bl.a. i:

- *Norsk språkbank* – utredning om et nasjonalt korpus for språkteknologi, 1999
- *Handlingsplan for norsk språk og IKT*. Rapport lagt fram av Norsk språkråd, februar 2001.
- *Samling og tilgjengeliggjøring av norske språkteknologiressurser* – Prosjektgruppe oppnevnt av Kultur- og kirkedepartementet. Rapport, oktober 2002
- St.meld. nr. 48 (2002–2003): *Kulturpolitikk fram mot 2014*
- *Norsk i hundre!* Norsk som nasjonalspråk i globaliseringens tidsalder – et forslag til strategi, Språkrådet 2005

Selskapet Nordisk Språkteknologi Holding AS gikk konkurs i november 2003. På det tidspunkt hadde selskapet gjennom en årrekke bygget opp svært omfattende språkressurser på norsk, svensk og dansk. I 2005 tok Språkrådet initiativ til å fremskaffe mer detaljert informasjon om språkressursene enn det som til da var tilgjengelig. I samarbeid med bostyrer ble det derfor foretatt en gjennomgang av NSTs språkressurser. Til å utføre oppdraget kontaktet Språkrådet derfor Avdeling for kultur, språk og informasjonsteknologi (Aksis) ved UNIFOB/ Universitetet i Bergen, som påtok seg oppdraget og en rapport ble levert til Språkrådet i desember 2005.[1] Rapporten la vekt på å dokumentere hva som finnes av språklige primærressurser. Dette omfatter akustiske databaser for talegjenkjenning og talesyntese, leksikalske databaser, tekstkorpus, dataverktøy for bearbeiding av ressursene, og dokumentasjon. Rapporten beskriver ressursenes plassering, omfang, format og kvalitet. For hver enkelt ressurskategori ble det vurdert om ressursene samsvarer med internasjonale standarder for språkressurser, slik de er definert av organisasjonen ELRA. Rapporten konkluderte med at dette materialet er verdifullt og må ivaretas og gjøres tilgjengelig for fremtidige forsknings- og teknologiutviklingsformål.

I 2006 ble det igangsatt forhandlinger med bostyret v/bostyrer Arne Laastad om et offentlig frikjøp av språkressursene, og en endelig avtale ble undertegnet 5. januar 2007. Et samarbeid mellom Norsk Språkråd, Universitetet i Bergen, Universitet i Oslo, NTNU i Trondheim og IBM Norge gjorde kjøpet mulig, Finansieringen ble delt mellom institusjonene slik:

Institusjon	**Andel (1000 kr)**
UiO	840
UiB	630
NTNU	630
Norsk Språkråd	400
Sum	**2 500**

[1] Gjennomgang og evaluering av språkressurser fra NSTs konkursbo, Rapport

I tillegg har IBM bidratt ved å gi avkall på sine immateriellrettslige tilgodehavender i boet. Intensjonen bak oppkjøpet er at det kjøpte materialet skal inngå som del av en fremtidig offentlig språkbank. Språkrådet har gjennom hele prosessen vært en pådriver i dette arbeidet og har tatt initiativ til en rekke politiske samtaler om språkbanken, bl.a. på statsrådsnivå.

Det formelle eierskapet forvaltes nå av en styringsgruppe bestående av Sverre Spildo (UiB), Arne Laukholm (UiO), Julie Feilberg (NTNU), Sylfest Lomheim (Språkrådet), og Roar Fundingsrud (IBM Norge). Styringsgruppen har ansvaret for materialet fram til etablering av en norsk språkbank og skal se til at materialet blir forvaltet og utviklet til beste for næringslivet, forskingen og det offentlige. Styringsgruppen arbeider også for at det offentlige etablerer og finansierer en norsk språkbank. Styringsgruppen har blant annet levert et høringsnotat om språkbanken knyttet til IKT-meldingen. Initiativet er omtalt i selve meldingen:

> Dei norske språkteknologiske fagmiljøa (universiteta, Forskingsrådet og Språkrådet) står samla bak eit uttalt behov og ønske om at ein nasjonal språkbank blir etablert.

2.1 SITS-prosjektets deltakere

Konsortiet som søker midler fra KUNSTI består av de samme aktørene som er med i gruppen som kjøpte NST-ressursene fra boet. Unifobs Avdeling for språk, kultur og informasjonsteknologi (Aksis) fungerer nå som vert for språkressursene fra NST og står som søkerorganisasjon og utførende enhet. Prosjektets konsortium består av følgende personer:

UiB/Unifob:	Gisle Andersen (prosjektleder), Koenraad de Smedt
UiO:	Hanne Gram Simonsen
NTNU	Torbjørn Svendsen
IBM	Roar Fundingsrud
Språkrådet	Torbjørg Breivik

3 Prosjektets faglige innhold, målsettinger og gjennomføring

Styringsgruppen har besluttet at språkressursene foreløpig plasseres ved Universitetet i Bergen, hvor ansvaret ivaretas av Aksis under faglig og administrativ ledelse av avdelingens forskningsdirektør, Gisle Andersen.

Materialet består av 15 servere, et CD-arkiv, et arkiv med DLT streamer tapes, og utskrevne dokumenter organisert i mapper. Alt materialet er fraktet til Universitetet i Bergen, og alt serverinnhold er kopiert, men det er ikke foretatt noen gjennomgang av dette, utover det som ble foretatt i 2005. Materialet er ennå ikke bearbeidet, modifisert eller omorganisert på noen måte. Et viktig moment er at CD-arkivet med ca. 2700 CD-plater inneholder blant annet unike data som ikke finnes i noe annet lagringsformat. Det er derfor et umiddelbart behov for å få kopiert og sikret disse dataene. En annen viktig oppgave er å overføre DLT-tapes og finne frem til seneste versjon av dataene.

For at den investering som er gjort skal ha noen verdi, er det nødvendig at dataene blir tilstrekkelig sikret, systematisert og dokumentert. SITS-prosjektet har som mål å utføre arbeidsoppgaver som er umiddelbart nødvendig for å sikre en forsvarlig forvaltning av disse ressursene. Dataene må være organisert på en slik måte at det ikke er noen risiko for tap av data, at det er enkelt å finne fram til bestemte deler av databasen og distribuere den til brukere, at tilstrekkelig dokumentasjon er tilgjengelig for eierne og brukerne, og at deskriptive metadata er i henhold til internasjonale standarder. Det overordnete målet er altså å systematisere og strukturere dataene slik at de fremstår som en velorganisert og oversiktlig

infrastruktur for forskning og utvikling, og å sørge for at brukermiljøene får umiddelbar nytte av dem. Dette innebærer følgende delmål:

- sikring av dataene, kopiering, lagring, sikkerhetskopiering (jf. 3.1)
- identifikasjon av hvor stor del av dataene som skal/kan gjøres tilgjengelig, fjerning av irrelevant materiell (jf. 3.2)
- systematisering, fjerning av duplikater, versjonskontroll (jf. 3.2)
- system for gjenfinning av data (jf. 3.2)
- strukturering av data, metainformasjon i henhold til internasjonal standard (jf. 3.2)
- avtaleverk for lisensiering til brukere (jf. 3.3)
- nettsted med innholdstjenester, søkbar ressurskatalog, dataprøver (jf. 3.4)
- informasjons- og formidlingsoppgaver, informasjonsmøter osv. (jf. 3.4)

Nedenfor gis en kort beskrivelse av hvert av disse delmålene og de oppgaver som vil bli utført som del av prosjektet.

3.1 Sikring av data

Som nevnt er allerede innholdet på serverne kopiert til UiBs servere. Serverdataene blir sikkerhetskopiert i henhold til UiBs rutiner, og duplikater fins på tre ulike lokasjoner i Bergen. Det er behov for å gjennomgå servernes innhold for å identifisere de deler som er relevante for en norsk språkbank og fjerne materiale som ikke skal inngå i språkbanken, slik som korrespondanse, e-post, medarbeidernes personlige kataloger osv. Dette vil bli foretatt, og utgangspunktet for arbeidet vil være evalueringsrapporten fra 2005.

I prosessen med å anskaffe dataene og overføre dem til Aksis i Bergen har det kommet frem at det finnes unike data i CD-arkivet, dvs. primærdata som ikke finnes i noen annen lagringsform. Det er derfor svært viktig at disse dataene kopieres til server, slik at de trygges for ettertiden og får en forsvarlig infrastruktur for sikkerhetskopiering. Dette arbeidet vil bli foretatt på en mest mulig tidseffektiv måte ved bruk av diskstasjon med kapasitet til samtidig kopiering av flere CD-er.

Det som finnes av DLT streamer tapes vil også bli overført til server. Til dette arbeidet trengs en ARC-server. Slike tapes ble brukt til NST for inkrementell sikkerhetskopiering. Innholdet på tapene må gjennomgås slik at man finner frem til seneste versjon og sørger for å merke tidligere versjoner som sådan. Det gjøres oppmerksom på at ingen språkdata vil bli endelig slettet, men en kopi av det opprinnelig anskaffete materialet vil bli spart for ettertiden.

3.1.1 Fremdrift og prosjektresultater (Sikring av data)

2007-07-15	Prosjektstart, oppstartsmøte, spesifikasjon av umiddelbare oppgaver
2007-07-15	Innkjøp og installasjon av utstyr for kopiering av DLT-tapes og CD-arkiv
2007-07-15	Igangsetting av kopiering av DLT-tapes og CD-arkiv
2007-08-15	Spesifikasjon av prosjektets arbeidsrutiner og ansvarsområder (notat)
2007-08-15	Spesifikasjon av språkbankens rutiner for sikkerhetskopiering (notat)
2007-10-01	Kopiering av DLT-tapes ferdigstilt
2007-10-15	Midtveisrapport til Styringsgruppe
2007-11-01	Kopiering av CD-arkiv ferdigstilt
2007-12-31	Sikring av data ferdigstilt
2007-12-31	Sluttrapport til Styringsgruppe

Milepæl: Utarbeidet detaljert plan for sikring av data – 15. august 2007

Milepæl: Oppnådd ansvarlig sikring av alle data – 31. desember 2007

3.2 Systematisering og strukturering av data

For systematisering og gjenfinning av data skal den teknologi som ble benyttet på NST videreføres og baseres på oppdaterte/moderniserte versjoner av denne. På denne måte oppnås ikke bare en effektiv måte å videreutvikle selve språkmaterialet fra NST, men også at dette skjer på en teknologisk plattform som skaper kontinuitet. Informasjonen skal systematiseres slik at man oppnår en korrekt og tidsmessig akseptabel tilgang til informasjon.

Det kreves en innsats for å finne frem til nøyaktig de filer som er relevante for leveranser til brukerne. Et utgangspunkt for dette arbeidet foreligger i form av beskrivelsene i 2005-rapporten, men det er likevel behov for å gjennomgå hele datasettet med hensyn til fjerning av duplikater og versjonskontroll. Dette vil bli gjort på en mest mulig automatisert og tidseffektiv måte. Arbeidet med fjerning av irrelevant materiell vil måtte gjøres manuelt.

Per i dag fremstår ikke dataene som en helhetlig infrastruktur men som en stor og uoversiktlig mengde med datafiler og dokumentasjon som man kun kan få tilgang til ved å lete/søke i servernes katalogstruktur. Prosjektet vil organisere dataene ved hjelp av et Content Management-system. Valg av programvare til dette formålet vil bli foretatt ved prosjektstart. Åpne løsninger som for eksempel IBMs løsninger (databaseteknologien DB2 og komponenter i WebSphere Voice Server) vil bli vurdert.

Innføring av et slikt system krever spesifikasjon av relevante kategorier og metadata, gruppering av data slik at både leverandør og bruker har oversiktlig tilgang til data som naturlig hører sammen (f.eks. alle norske akustiske opptak knyttet til en bestemt leveranse), annotering av metainformasjon og implementering av funksjonalitet for søking og leting i dataene. Systemet må omfatte all tilgjengelig metainformasjon om språkressursene, slik som datatype (akustisk database, leksikon, tekst), primærformål for anskaffelse (talegjenkjenning, talesyntese, språkmodellering), filformat (pcm/waw, txt/doc osv), størrelse (ordtilfang, antall ord, antall opptak, filstørrelse), dato for innsamling, geografisk område osv. Denne delen av arbeidet vil ta utgangspunkt i internasjonale standarder for språkressurser (ELRA, CLARIN).

3.2.1 Fremdrift og prosjektresultater (Systematisering og strukturering av data)

2007-08-15	Teknisk spesifikasjon for kategorisering, metainformasjon og programvare
2007-09-01	Gjennomgang av servere, merking av relevant/irrelevant materiale
2007-09-01	Content Management-system implementert
2007-10-01	Gjennomgang av servere, versjonskontroll
2007-10-15	Midtveisrapport til Styringsgruppe
2007-11-01	Content Management-system fullt operativt
2007-12-31	Content Management-system, språklige data og dokumentasjon lagt inn
2007-12-31	Sluttrapport til Styringsgruppe

Milepæl: Content Management-system oppdatert med innhold – 31. desember 2007

3.3 Avtaleverk

Det er under utarbeidelse et avtaleverk som skal ivareta opphavsrettigheter, bruksrettigheter og tilgjengelighet for de språkressursene gruppen har kjøpt. Avtaleverket må være på plass før man kan stille ressursene til disposisjon for aktuelle brukere. Det internasjonale avtaleverket som ELDA legger til grunn for sin virksomhet vil, så langt det er mulig, danne grunnlag for de norske avtalene. ELDA vil også være en sentral samarbeidspart i dette arbeidet. De norske

avtalene vil ikke gi mulighet for noen eksklusive rettigheter til hele eller deler av materialet for kortere eller lengre tid for aktuelle brukere. Det vil bli utarbeidet to standardavtaler: en som brukes til forskningsmiljøene og en som brukes til private aktører i næringssammenheng. Målet er å gjøre språkressursene tilgjengelig for brukerne innenfor forskning og industri så raskt som mulig. Avtalene vil bli kvalitetssjekket av jurister med kompetanse på avtaleverk som omhandler immaterielle verdier.

3.3.1 Fremdrift og prosjektresultater (Avtaleverk)

2007-08-15	Møte i juridisk gruppe, oppsummering av status, videre planlegging
2007-09-01	Kartlegging og analyse av internasjonale avtaleverk
2007-10-01	Utkast til avtaletekster
2007-10-15	Midtveisrapport til Styringsgruppe
2007-11-01	Ferdigstilling av avtaletekster
2007-12-31	Evaluering
2007-12-31	Sluttrapport til Styringsgruppe

Milepæl: Utarbeidet sett av avtaletekster til ulike brukere – 1. november 2007

3.4 Nettportal, informasjon og formidling

En betydelig del av prosjektinnsatsen vil bli viet formidling av prosjektet, i vid forstand. Det viktigste formidlingstiltaket er nettportal med innholdstjenester. Denne vil først og fremst ha en ekstern funksjon og vil være brukernes viktigste kilde til informasjon om språkressursene, det spesifikke SITS-prosjektet og om Språkbanken for øvrig.

Portalen vil inneholde all tilgjengelig metainformasjon om språkressursene (jf. 3.2). Denne informasjonen vil bli fremskaffet først på grunnlag av opplysninger i 2005-evalueringen, og ressursene vil bli kategorisert i henhold til klassifiseringen som fremkommer her. Senere vil det bli supplert med informasjon fra øvrig dokumentasjon som foreligger fra NST-tiden, og til sist vil informasjon bli sjekket etter inspeksjon av selve dataene. Ressursoversikten vil være etter mønster fra den europeiske organisasjonen ELRAs tilsvarende oversikt.[2] Etter hvert vil det bli viktig å samordne kategoriseringen med arbeidet som gjøres innenfor CLARIN-prosjektet.[3] Dette ESFRI-prosjektet, som NFR har gitt sin støtte til, vil imidlertid ikke starte opp før tidligst i januar 2008.

All denne informasjonen vil først og fremst bli tilgjengelig på norsk, fordi norske og nordiske brukermiljøer er de primære målgruppene. I den grad det er kapasitet til det innenfor den korte og intensive prosjektperioden, vil viktig informasjon også bli lagt ut på engelsk. På lenger sikt (etter endt prosjektperiode) vil målet være å ha fullt operative innholdstjenester også på engelsk.

Ressursoversikten skal være tilgjengelig på en slik måte at brukerne kan finne frem til en ressurs både ved å søke og ved å klikke (browse) seg gjennom de ulike delene. Man skal også kunne sortere ressursoversikten i henhold til den ulike metainformasjonen som foreligger. Portalen vil også inneholde informasjon for brukerne om retningslinjer for bruk, juridiske begrensinger (begrenset til forskningsformål/kommersielle formål) og hvordan man skal forholde seg for å anskaffe en bestemt ressurs.

[2] http://catalog.elra.info/
[3] CLARIN – Common Language Resources and Technology Infrastructure; http://www.clarin.eu/

Det skal finnes dataprøver tilgjengelig, dog er det nødvendig å ivareta hensynet til personvern og andre bruksbetingelser knyttet til dataene og ellers vise aktsomhet ved publisering av dataprøver (en bidragsyter til databasen skal ikke risikere å kunne klikke seg frem til et lydopptak av seg selv). Det vil bli praktisert en åpen linje når det gjelder tilgjengeliggjøring av dokumentasjon. Det som fins av relevant ressursdokumentasjon fra NST-tiden vil bli publisert, og ny dokumentasjon vil bli fremstilt og gjort tilgjengelig. Portalen vil også inneholde prosjektspesifikke nyheter, med korte meldinger om arbeidets fremdrift. Dette vil også være en portal for nyheter om språkbanksakens fremdrift, men mer allmenne språkteknologinyheter vil ikke være fokus her, da man kan henvise til den eksisterende språkteknologiportalen Norsk dokumentasjonssenter for språkteknologi.[4]

Formidlingsarbeidet omfatter også andre tiltak. Det vil bli holdt 1-2 informasjonsmøter med brukermiljøene. Det vil bli skrevet en teknisk rapport knyttet til de løsninger som er valgt for teknisk infrastruktur, metainformasjon osv. En intern instruks for formidling, klargjøring og forsendelse av data vil foreligge tidlig i prosjektfasen (medio august). Videre vil det være prosjektleders oppgave å supplere informasjon til Styringsgruppen, til bruk som «ammunisjon» i den videre politiske prosessen frem mot en endelig politisk språkbankvedtak, samt å bistå Styringsgruppen i kontakten med media og offentlighet, når dette finnes naturlig.

3.4.1 Fremdrift og prosjektresultater (Nettportal, informasjon og formidling)

2007-08-15	Planlegging av nettportalens utforming og innhold
2007-09-15	Pilotversjon av nettportal klar
2007-09-15	Ressursoversikt klar i første versjon, deretter kontinuerlig oppdatert
2007-10-01	Nasjonalt møte med brukermiljøene
2007-10-15	Midtveisrapport til Styringsgruppe
2007-12-01	Nettportal klar i fullt operativ versjon
2007-12-01	Ressursoversikt klar
2007-12-31	Evaluering av portalens utforming og innhold
2007-12-31	Nytt møte med brukermiljøene, hvis behov
2007-12-31	Sluttrapport til Styringsgruppe

Milepæl:	Ferdigstilt og fullt operativ nettportal – 1. desember 2007
Milepæl:	Ferdigstilt ressursoversikt

4 Konkrete planer for bruk av ressursene til forskning og næringsutvikling

Hvert av de deltakende universitetene har konkrete planer for bruk av språkressursene, som inngår som en betydelig komponent i miljøenes forskningsstrategier. Dels er dette arbeidet allerede i ferd med å komme i gang. Fagmiljøene understreker betydningen av å ha slike ressurser tilgjengelig. Det har også kommet mange henvendelser fra kommersielle aktører som er interessert i å ta materialet i bruk. Nedenfor konkretiserer vi planene hos enkelte aktører innen akademia og næringsliv. Det understrekes at de nevnte prosjektene ikke inngår i selve SITS-prosjektet og søkes ikke finansiert via KUNSTI-programmet.

4.1 Akademisk forskning

4.1.1 NTNU

Språkteknologiforskningen ved NTNU er et tverrfaglig samarbeid mellom Institutt for språk og kommunikasjonsstudier, Institutt for datateknikk og informasjonsvitenskap og Institutt for

[4] http://www.norskdok.uib.no/

elektronikk og telekommunikasjon. Alle disse miljøene har stort behov for språkressursene som ligger i NST-databasen i sitt forskningsarbeid. Spesielt gjelder dette taleteknologi-forskningen. Store mengder relevante akustiske taleopptak og tekstsamlinger er nødvendig for å bygge statistiske modeller for tale og språk. Mangel på tilstrekkelige mengder med norske språkressurser har vært et hinder for forskning og utvikling av avansert norsk talegjenkjenning. Med tilgang til NST-databasen er NTNU allerede i gang med å utvikle en norsk storvokabular talegjenkjenner som vil være en basis for videre forskning innen dette området, og som vil være en viktig komponent i anvendelser som dikteringssystemer og informasjonsgjenfinning i multimedia-databaser. Forskningen innen talesyntese vil også dra stor nytte av dataressursene. Akustiske data er av stor interesse for fonetikkmiljøet, og tekstsamlinger og leksikalske data er nødvendige for fagmiljøene innen lingvistikk og naturlig-språkprosessering. I tillegg til de språkteknologiske miljøene vil tilgang til norske språkdata være av stor betydning for de språkvitenskapelige miljøene.

4.1.2 Universitetet i Bergen

Universitetet i Bergen har flere språkfaglige miljøer som vil være interessert i NST-databasene som forskningsressurs, bl.a. Seksjon for lingvistiske fag, Nordisk institutt og Unifobs Avdeling for kultur, språk og informasjonsteknologi. De leksikalske databasene betraktes som interessant og som et godt grunnlag for videreutvikling og tilpassing av norske leksikalske databaser, gjerne flettet sammen eksisterende databaser som SCARRIE og LEXIN. Et aktuelt forskningstema er ulike aspekter knyttet til ortofoni, samsvaret mellom ortografi og uttale i norsk. Materialet kan være et utgangspunkt for videre bearbeiding og utvidelse i retning av en Norsk versjon av CELEX. De akustiske ressursene betraktes som interessante for studier av fonetisk variasjon i norsk, mens tekstressursene er aktuelle som supplement til eksisterende ressurser som Norsk aviskorpus.

4.1.3 Universitetet i Oslo

Ved Universitetet i Oslo er det flere aktuelle forskningsmiljøer som vil dra nytte av språkressursene, spesielt for språkteknologiske formål, men også mer generelt for lingvistisk og fonetisk forskning. Det gjelder f.eks. Tekstlaboratoriet ved ILN, Enhet for digital dokumentasjon, Forskningsgruppe for logikk og naturlige språk ved Institutt for informatikk, USIT, i tillegg til språkvitenskapelige og fonetiske miljøer ved Institutt for lingvistiske og nordiske studier. De leksikalske databasene vil være interessante i den formen de finnes her med transkripsjon og ev. nye annoteringer. Det som fins av juridisk tilgjengelig talespråksmateriale vil være interessant i talemålsforskning og dialektforskning – det at det også er nordiske ressurser her, øker interessefeltet. USIT er spesielt interessert i de akustiske ressursene for å utvikle syntetisk tale på web.

4.2 Næringsutvikling

Tilgang på språkdata er en forutsetning for utvikling eller tilpasning av norsk språkteknologi. For et lite land som Norge er det i tillegg av stor betydning at grunnleggende språkressurser er billige eller gratis. For anvendelser som talegjenkjenning er den grunnleggende teknologien i hovedsak felles for mange språk. For eksempel kan det nevnes at Microsoft Vista leveres med integrert talegjenkjenning for noen utvalgte språk. Norsk ligger per i dag langt nede på Microsofts prioriteringsliste, noe som kan endre seg når en godt tilrettelagt norsk språkdatabase blir tilgjengelig. I tillegg til store internasjonale aktører er språkressursene av stor betydning for mindre språkteknologibedrifter.

Det kommer stadig henvendelser fra språkteknologibedrifter i inn- og utland om tilgang til ressursene. Dette understreker behovet for å få i stand en systematisert tilgjengeliggjøring. Et

representativt eksempel som kan trekkes frem er et initiativ fra Bredtvet kompetansesenter. Deres erfaring tilsier at overgangen til elektronisk informasjon ikke nødvendigvis gjør tekst mer tilgjengelig for personer med manglende funksjonell lesekompetanse. Deres leseproblemer overføres fra papirbaserte dokumenter, men teknologien gir dem helt unike muligheter til å benytte talesyntese for å få opplest tekst slik at de kan forstå innholdet og dermed tilegne seg informasjon og kunnskap. Talesyntese sees primært på som lesehjelpemiddel, men gir også støtte i skriveprosessen, mens talegjenkjenning som skrivehjelpemiddel vil kunne gjøre skriveprosessen enklere og raskere når de kan bruke stemmestyring. Slike løsninger vil også kunne avhjelpe skrivevansker som bevegelseshemmede eller personer med musesyke opplever. I skolesammenheng vil en god kvalitet på talesyntesen gi økt tilgang til fagstoff og åpne kulturelle verdier for både lesesvake og for personer med norsk som andrespråk. God kvalitet i form av lett oppfattbar talesyntese vil være viktig også for eldre personer og for personer med kognitiv svikt. På denne bakgrunn har senteret tatt initiativ til et utviklingssamarbeid med prosjektgruppen som baserer seg på bruk av ressurser i NST-materialet.

4.2.1 IBM

Gjennom flere år har utvikling av en basisteknologi for å støtte språkbaserte løsninger (tekst til tale og tale til tekst) vært et strategisk satsningsområde for IBM. Ved IBMs internasjonale laboratorier som arbeider med utvikling av denne type løsninger, finnes det nå også prosjekter som omfatter tale til tale (flerspråklig og uten tolk). Videreføring av den teknologi som ble utviklet ved NST og IBMs engasjement i dette prosjektet, og etablering av en norsk språkbank, er derfor av stor strategisk betydning for IBM. Med teknologien kalt WebSphere har IBM etablert en teknologisk plattform som bygger på åpne standarder og i sin helhet på en serviceorientert arkitektur (SOA). Dette er en plattform som kan videreutvikles og tilpasses de språkteknologiske løsninger man ønsker å utvikle/teste ut. IBM ser det videre som strategisk viktig å få etablert en Språkbank for det nordiske markedet og bidra til at det etableres et fornyet samarbeid mellom IBMs internasjonale nettverk av laboratorier og Språkbanken for forskning og utprøving av ny teknologi.

5 Prosjektorganisering

Det tekniske og organisatoriske arbeidet vil i hovedsak bli utført ved Aksis. Prosjektleder er Forskningsdirektør/dr.art. Gisle Andersen, som har erfaring både som utvikler og språkteknolog fra NST og fra arbeidet med evalueringen på oppdrag fra Språkrådet. Prosjektgruppen består av programmerere og fagkonsulenter ved Aksis.

Med i prosjektet bidrar også to faggrupper som har viktige strategiske, operative og rådgivende funksjoner. En teknisk arbeidsgruppe er satt sammen av representanter fra eierne og har ansvar for å fatte beslutninger om tekniske løsninger for systematisering av data, teknisk infrastruktur osv. Gruppen ledes av Roar Fundingsrud (IBM) og består ellers av Dag Jacobsen, Lars Fanebost og Frode Tveit (IBM), Morten Erlandsen (UiO), Torbjørn Svendsen (NTNU), Torbjørg Breivik (Språkrådet), Sindre Sørensen og Gisle Andersen (Aksis/UiB). En juridisk arbeidsgruppe har ansvar for kartlegging av og veiledning om juridiske forhold knyttet til videreformidling av språkressurser, samt utforming av avtaleverk. Gruppen består av rådgiver Torbjørg Breivik (Språkrådet, leder), juridisk rådgiver Kitty Amlie Tverrå (UiB) og språkingeniør Kristin Hagen (UiO).

Prosjektleder samarbeider nært med disse to faggruppene, og mottar innspill og beslutninger fra dem, som han har ansvar for å iverksette. Prosjektleder rapporterer løpende om arbeidets

fremdrift til Styringsgruppen for språkbanken. Det vil bli holdt månedlige møter i de to faggruppene gjennom prosjektperioden, som varer fra 15. juli 2007 til 31. desember 2007.

6 Budsjett

Det understrekes at det som søkes finansiert gjennom KUNSTI er kostnadene knyttet til *etablering av en nasjonal forskningsinfrastruktur* basert på de eksisterende språkressursene. Mer rutinepregete og gjentakende driftsoppgaver søkes ikke finansiert.

Det legges opp til en kort *planleggingsfase* (juli-august 2007) og en *implementeringsfase* (september-desember 2007) som inngår i selve SITS-prosjektet, før man starter en fullt operativ driftsfase fra januar 2008. Finansieringen av driftsfasen er eiernes ansvar og dette ansvaret vil bli ivaretatt av Styringsgruppen på en slik måte at Språkbanken sikres kontinuerlig drift (ev. på et minimumsnivå i mangel av ekstern finansiering) i interimperioden mellom det avsluttede SITS-prosjektet og frem mot en formelt etablert norsk språkbank.

På lengre sikt er det nødvendig med et mer omfattende arbeid med språkressursene enn det som det er mulig å få til i løpet av SITS-prosjektet, som kun løper ut 2007. Dette dreier seg først og fremst om oppdatering og fornying av ressursene og harmonisering av dem i henhold til internasjonale standarder. Dette vil bli søkt finansiert gjennom CLARIN-prosjektet. Det gjøres oppmerksom på at prosjektleder Gisle Andersen også er nasjonal koordinator for CLARIN-samarbeidet. Dessuten arbeider Styringsgruppen kontinuerlig for å få til vedtak om en fast statlig bevilling til etablering og drift av en norsk språkbank.

I dette prosjektet søkes det om midler til dekking av arbeidskostnader, programvare og maskinvare, reiser og møtevirksomhet og et minimum av generelle driftskostnader. Budsjettet er som vist i tabellen.

Timebasert lønn	Prosjektleder GA, programmerer AK, programmerer SS; fagkonsulent AKL, fagkonsulent KH, fagkonsulent PV	793 000
Programvare og maskinvare		80 000
Reiser/møter	Teknisk/juridisk faggruppe, styringsgruppe, spesifikke tekniske møter, møter med brukermiljøene; totalt 10 nasjonale møter m. i snitt 6 deltakere	120 000
Driftskostnader		10 000
Totalt		**1 003 000**

7 Referanser

Andersen, Gisle. Gjennomgang og evaluering av språkressurser fra NSTs konkursbo. Rapport utarbeidet på oppdrag fra Språkrådet, 2005.

European Strategy Forum on Research Infrastructures (ESFRI): European Roadmap for Research Infrastructures, Report 2006.

Handlingsplan for norsk språk og IKT. Oslo: Norsk språkråd, 2001.

Norsk i hundre! Norsk som nasjonalspråk i globaliseringens tidsalder – et forslag til strategi. Oslo: Språkrådet, 2005.

Norsk språkbank – utredning om et nasjonalt korpus for språkteknologi, 1999

Samling og tilgjengeliggjøring av norske språkteknologiressurser. Oslo: Kultur. og kirkedepartementet, 2002.

Stortingsmelding nr. 48 (2002–2003): Kulturpolitikk fram mot 2014.