Språkrådet

PLAN FOR ETABLERING AV NORSK SPRÅKBANK

15. august 2008

Utarbeidet av en arbeidsgruppe nedsatt av Språkrådet

Forord

Regjeringa har i to stortingsmeldinger omtalt etablering av en norsk språkbank som det viktigste tiltaket for å bevare og styrke norsk språk i en tid med globalisering og domenetap. I den siste, Mål og meining, som kom i juni 2008, slår Regjeringa fast at Språkbanken skal etableres (St.meld. nr. 35 (2007-2008), pkt. 7.5.6.2, s. 136) fra 1. januar 2009.

Språkrådet fikk i brev 6. mai 2008 følgende oppdrag fra Kultur- og kirkedepartementet: "Vi ber om at Språkrådet nedsetter en arbeidsgruppe som innen 15. august 2008 skisserer en plan for hvordan språkbanken skal etableres. Planen skal angi hvilke økonomiske og personellmessige ressurser språkbanken vil kreve det første driftsåret, og hvordan språkbanken skal organiseres. Arbeidsgruppen skal også skissere en realistisk strategi og opptrappingsplan for språkbanken i årene 2010 – 2014."

Arbeidsgruppen som ble nedsatt av Språkrådet 15. mai 2008, leverer med dette planen for etablering av Norsk språkbank. Arbeidsgruppen står samlet bak planen, og det har vært kontakt med Kultur- og kirkedepartementet, Kunnskapsdepartementet, Fornyings- og administrasjonsdepartementet og Nærings- og handelsdepartementet undervegs i arbeidet for å avklare sentrale problemstillinger. Konklusjonene står for gruppens oppfatninger.

professor Torbjørn Svendsen, NTNU og fagmiljøene,
assisterende direktør Sverre Spildo, UiB og eierne av NST-ressursene,
førsteamanuensis Jan Olav Fretland, Høgskolen i Sogn og Fjordane og styret i Språkrådet
seniorrådgiver Torbjørg Breivik, Språkrådet.

Advokat Kristine M. Madsen, Bull & Co Advokatfirma AS, har vært gruppens juridiske rådgiver.

Visjon

Språkbanken skal være det naturlige samlingspunktet for lagring og distribusjon av offentlige og private digitale språkressurser.

Målet for Språkbanken er:

- **å være infrastruktur i språkteknologisk forskning, utvikling og produkt- og tjenestetilpasning for norsk språk**

Innhold

0 Sammendrag og konklusjoner

Sammendrag

Arbeidsgruppen har lagt til grunn at Norsk språkbank skal etableres for å bidra til å realisere formålet i St.meld. nr. 35 om en ny, strategisk språkpolitikk med helhetsperspektiv på språk og samfunn, en politikk som skal sikre det norske språkets posisjon som et fullverdig, samfunnsbærende språk i Norge. Videre har arbeidsgruppen lagt vekt på stortingsmeldingas prinsipp om at språkpolitikken skal være sektorovergripende med kulturpolitisk forankring.

Arbeidsgruppen har basert arbeidet på rapporten "Samling og tilgjengeleggjering av norske språkteknologiske ressursar" fra 2002 med vedlegg. Vurderingene i rapporten er gjennomgått og vurdert på nytt. Oversikten over hvilke språkressurser som befinner seg hvor, er oppdatert, men gruppen anbefaler at den gjennomgås mer detaljert og at aktualiteten vurderes bl.a. med hensyn til rettigheter til videre bruk av de enkelte ressursene. I tillegg til ressursene som omtales i rapporten fra 2002, har vi i dette arbeidet sett på verktøy for bearbeiding av språkressurser for videre bruk. Grunnen er at det særlig på dette området har skjedd mye i perioden 2002-2008.

Internasjonalt er det utarbeidet en standardmal for hva som bør inngå i en grunnleggende språkteknologisk ressurssamling for at den skal være formålstjenlig for språkteknologisk bruk. Standarden kalles "Basic LAnguage Resource Kit", og det er utarbeidet en foreløpig norsk versjon av denne standarden – en norsk BLARK. Standarden bør legges til grunn for det videre arbeidet med innsamling av ressurser og verktøy, slik at den gir føringer når det gjelder rekkefølge og prioritering av innhold. Eksisterende språkressurser og –verktøy er omtalt og vurdert i den grad det har vært mulig. Det må kommenteres at det ikke har vært tid til en grundig gjennomgang og vurdering av den enkelte ressurs.

En språkbank er en kostbar investering for ethvert språksamfunn, og kostnadene varierer lite fra språk til språk. For norsk vil man måtte legge til en del fordi vi har to likestilte målformer som begge skal representeres i Språkbanken, og materiale for begge målformene må bearbeides og tilrettelegges for gjenbruk. Kalkylene over investeringer og administrative kostnader tar utgangspunkt i beregningene fra 2002, men er justert ut fra pris- og lønnsøkningen som har vært i perioden, og det er tatt hensyn til at den teknologiske utviklinga har gått framover. Tilgang til godt merkede og tilrettelagte språkressurser etter et norsk språkteknologiforetak (NST) som gikk konkurs i 2003, er tatt hensyn til i disse kalkylene.

Konklusjoner

Norsk språkbank skal være et tiltak for å oppnå hovedmålet med stortingsmelding 35 (2007–08): "... språkpolitikk med det overordna målet å sikra det norske språkets status og bruk på alle samfunnsområde, slik at norsk kan bestå som eit fullverdig, samfunnsberande språk". Språkmeldinga setter også klare mål for en norsk språkbank: "Språkbanken skal vera ein stor, samla, nasjonal språkressurs som er kvalitetssikra og bygd opp etter internasjonale standardar."

Vår visjon for Språkbanken er at den skal være det naturlige samlingspunktet for lagring og distribusjon av offentlige og private digitale språkressurser. Ut fra dette foreslår arbeidsgruppen:

- Norsk språkbank blir etablert som aksjeselskap med eget styre fra 1.1.2009.

- Staten ved Kultur- og kirkedepartementet blir hovedeier i selskapet.

- En detaljert norsk BLARK-analyse legges til grunn for arbeidet med etableringen.

- Etableringsfasen for Norsk språkbank settes til 6 år, med en investeringsramme på 90 mill. Årlig administrativ ressurs er beregnet til 3 mill.

- Språkbankens formål er å lette tilgangen til eksisterende språkressurser og verktøy for språkteknologisk forskning og utvikling for både private og offentlige aktører. En forutsetning er også at de aktuelle miljøene bidrar til utvikling av Språkbankens ressurser.

- Språkbanken vil ha en liten administrasjon og bør lokaliseres i tilknytning til et relevant språkteknologisk eller språkfaglig miljø. Arbeidsgruppen mener at en samlokalisering med Språkrådet kan være en god løsning. Det kan likevel tenkes andre løsninger for lokalisering, ut fra arbeidsgruppens prioriteringer.

- Språkbanksatsingen må følges opp og koordineres med en sterk satsing på språkteknologisk forskning.

- Arbeidsgruppen har diskutert sterke og svake sider ved den norske språkbanksatsingen i en egen analyse som er vedlagt rapporten (SVOT-analyse).

1 Innledning

Til grunn for arbeidet sitt legger arbeidsgruppen de overordnete mål og føringer som ligger i Stortingsmelding nr. 35. Det vil si at opprettelsen av en norsk språkbank skal være et ledd i målet om "at norsk skal vera hovudspråk og nasjonalspråk i Noreg, at norsk skal vera eit samfunnsberande og fullverdig språk, at det skal leggjast til rette for at nynorsk blir meir reelt likestilt med bokmål, at det offentlege skal leggja vinn på å føra eit korrekt og forståeleg språk, og at alle skal ha rett til språk, rett til nasjonalspråk, rett til morsmål og rett til å kunna læra seg framande språk" (s. 23).

I St.meld. nr. 48 (2002–2003) "Kulturpolitikk fram mot 2014" er det en fyldig omtale av en norsk språkbank og hvilke muligheter den gir for utvikling av forskning og næringsutvikling innenfor språkteknologi. Det blir også slått fast at "Språkteknologifeltet kan vera ein av dei framste arenaene der kampen om norsk språk og kultur vil utspela seg i tida framover."(s. 196). Det heter også at "Den teknologiske utviklinga har gjort det endå tydelegare enn før at det er på andre arenaer enn der kulturpolitikken tradisjonelt opererer, at viktige premissar for språkutviklinga vert forma." (s. 197).

St.meld. nr. 35 (2007–2008) "Mål og meining" går videre og slår fast at "Språkteknologi [er] med på å leggja til rette for auka demokratisk deltaking i samfunnet ved å gjera informasjon og tenester tilgjengelege for alle". Nye produkter og nye tjenester som baserer seg på språkteknologi eller har språkteknologiske komponenter, kommer stadig på markedet, og nye anvendelsesområder dukker stadig opp. Det er få av disse produktene som er tilrettelagt for norsk språk. Internasjonalt har man lenge arbeidet med tilrettelegging av denne typen samlinger for språkteknologiske formål, og nå slås det fast at "Regjeringa har som mål i dei næraste åra å få bygd opp ein norsk språkbank." (s. 133 ff.).

Kultur- og kirkedepartementet viser til at organisering og finansiering av Språkbanken må avklares med tre andre departementer: Kunnskapsdepartementet, Fornyings- og administrasjonsdepartementet og Nærings- og handelsdepartementet. Underveis i prosessen ble det arrangert et møte mellom arbeidsgruppen og saksbehandlere i berørte departementer for å hente inn synspunkter på sentrale spørsmål i utredningen.

Arbeidsgruppen ser Språkbanken som et statlig tiltak som skal utføre en viktig kulturpolitisk oppgave for Norge, og som har et sektorovergripende, samfunnsnyttig formål. I og med at Språkbanken skal tjene hele samfunnet, både det offentlige (forskningsformål) og det private (næringsutvikling), bør den være en selvstendig organisasjon med et uavhengig og bredt sammensatt styre.

Språkbankens primære formål er å gjøre språkressurser tilgjengelige for språkteknologiske formål. Det finnes også andre samlinger av språkressurser i Norge som bygges opp med andre primærformål. Dette vil ikke bety at det vil bli et konkurranseforhold om tilgang til ressurser som skal innlemmes i samlingene, spesielt ikke siden Språkbanken verken vil ha eksklusiv bruksrett eller eierskap til ressursene som skal distribueres. Det er snarere slik at

de ulike samlingene kan utfylle hverandre, slik at ressurser som er samlet inn, kan brukes også til andre formål enn de primære, og data og verktøy kan deles. Et godt eksempel er EU-initiativet CLARIN, der målet er å gjøre ulike typer tekster (eldre og nyere) elektronisk tilgjengelig for forskning innenfor humaniora. Norske forskningsinstitusjoner deltar i dette arbeidet, og deler av innholdet i Språkbanken vil kunne være en del av et norsk CLARIN.

Arbeidsgruppen vil også understreke betydningen av at det parallelt med selve opprettelsen av en språkbank blir opprettet nye forskningsprogram som kan utnytte Språkbanken til språkteknologisk forskning og utvikling, videreutvikle ressursene i Språkbanken og utvikle nye elementer som kan inngå i Språkbanken. Vi er glad for at det blir lagt stor vekt på at samarbeidet mellom flere departementer er nødvendig for å realisere intensjonene både for språkpolitikken generelt og for Språkbanken spesielt. Vi vil også understreke spesielt det språkmeldinga sier om at Språkbanken må være bygd opp etter internasjonale standarder.

Flere utredninger omtaler en norsk språkbank. Et hovedarbeid ligger i rapporten "Samling og tilgjengeleggjering av norske språkteknologiressursar" (Språkrådet 2002). Rapporten legger til grunn at om vi skal få norskspråklige språkteknologiske tjenester og produkter, bør man opprette en språkbank som er i statlig eie. Markedet er for lite til at kommersielle aktører vil påta seg kostnadene med å samle inn og tilrettelegge de nødvendige språklige råvarene.

Gruppen har i det konkrete arbeidet med rapporten tatt utgangspunkt i planen fra 2002. Det har kommet til andre og nye språkressurser etter den tid, og arbeidsmåtene er i noen grad forandret. Planen fra 2002 hadde f.eks. ikke en detaljert omtale av verktøy for bearbeiding av språkressurser som en del av innholdet i Språkbanken. Situasjonen i dag tilsier at verktøy for bearbeiding og tilrettelegging av språkressurser vil være en del av det som skal gjøres tilgjengelig gjennom Språkbanken. Dette er også i tråd med internasjonale føringer (se avsnittet om BLARK). Det er nødvendig med en grundigere kartlegging av hvilke språkressurser som finnes, enn den arbeidsgruppen har vært i stand til med den korte tiden vi har hatt til rådighet. En slik videre kartlegging og systematisering kan med fordel gjøres høsten 2008. Gruppen har lagt dagens behov til grunn for sin prioritering av hvor man bør starte.

Kostnadsvurderingene fra 2002 konkluderte med at det trengs totalt 100 mill.kr over fem år. Arbeidsgruppen har gått gjennom beregningene på nytt og konkluderer med at totalsummen til utvikling fortsatt kan settes til ca 90 mill.kr. Dette svarer til en utgiftsreduksjon på om lag 25 % før korreksjon for lønnsutviklingen. Utgifter til administrasjon av Språkbanken kommer i tillegg. Gruppen holder fast på prinsippene om en liten administrasjon og at arbeidet utføres i de relevante fagmiljøene.

Arbeidsgruppen har laget en analyse av hvilke utfordringer og hindringer vi ser for å kunne oppfylle målene for Språkbanken. Denne SVOT-analysen er vedlagt rapporten.

Arbeidsgruppen har vurdert organisasjonsformer som forvaltningsorgan, stiftelse og aksjeselskap, og vi har kommet til at aksjeselskap er best tjenlig for Språkbankens

formål. Ut fra dette har arbeidsgruppen også skrevet et utkast til vedtekter. Utkastet til vedtekter er vedlagt.

2 Eksisterende materiale

Arbeidsgruppen har ikke hatt anledning til en fullstendig kartlegging av eksisterende materiale, ikke minst fordi det har vært sommer og ferietid. Mange har levert oversikter, men her gjenstår en del arbeid før vi kan si at dette er en komplett oversikt. Det vi har nå, er likevel tilstrekkelig til å starte med. I kapittel 4.6 omtales forberedende oppgaver som er identifisert undervegs i dette arbeidet, og som vil være bra å få gjort i løpet av høsten 2008.

Det vil være tilstrekkelig med bruksrett til språkressurser for Språkbankens formål. Det er ikke nødvendig med eiendomsrett eller enerett til ressurser i Språkbanken. Den vil kun selge bruksretter til innholdet den disponerer. Vedlikehold og videreutvikling av innholdet i Språkbanken vil baseres på samarbeid med fagmiljøene.

2.1 Taledata

Talematerialet etter Nordisk Språkteknologi Holding AS (NST) er sikret for innlemming i Språkbanken. Det må avklares med Interimsstyret for språkbanken (konsortiet) hvordan overføringen til Språkbanken skal skje. Konsortiet som kjøpte boet etter NST, består av Universitetet i Bergen, Universitetet i Oslo, Norges teknisk-naturvitenskapelige universitet, IBM Norge og Språkrådet.

Alle rettigheter til taledelen av NST-materialet er avklart, materialet er gjennomgått og klargjort for gjenbruk. Materialet kan distribueres som det er, men filene bør konverteres til et format som tilfredstiller DIFIs retningslinjer for lagringsformat og internasjonale standarder på området. Materialet må gjennomgås for å avdekke mangler og for å se hvilke suppleringer som trengs. Det må bl.a. suppleres med nye data for å få en bedre fordeling når det gjelder alder. I tillegg mangler nynorsk materiale. Metainformasjonen bør oppdateres. Validering og kvalitetssikring av ressursene må til. Det betyr at vi alt har mye materiale av høy kvalitet for taledelen i Språkbanken, men det bør gjøres en del arbeid for å oppdatere det og tilpasse det dagens standarder.

En oppdatert oversikt over eksisterende taledata er gitt i Tabell B.1 i Vedlegg B. Tabell 3 i kapittel 4 inkluderer en sammenstilling av tilgjengelige taleressurser fordelt på typer. I tabellen er det også angitt tall for minimumsinnhold og ønsket innhold i Språkbanken.

2.2 Tekstdata

Mye tekstdata er tilgjengelig, og mengden tekstdata har økt betydelig siden 2002. Imidlertid er det et gjennomgående trekk at bruk av samlingene enten er avgrenset til forskningsformål, og/eller at bruksrett må avklares med opphavseierne. Dette gjelder for eksempel store tekstsamlinger som NST-tekstbasen og Norsk aviskorpus, som til sammen utgjør over en milliard løpende ord. Det trengs generelt en avklaring av rettigheter med de store tekstleverandørene. Det eksisterende tekstmaterialet er skjevt fordelt med en overvekt av aviser og sakprosa i materialet, og det trengs en gjennomgang for å innhente annet materiale slik at man får et innholdsmessig balansert tekstkorpus. Innholdet må også balanseres med hensyn til de to målformene, og det er særlig materiale på nynorsk som kan bli vanskelig å få tilstrekkelig mengde av og stor nok variasjon av med hensyn til typer tekst.

Det er etablert flere samlinger med flerspråklig parallelltekst. Mye av dette er imidlertid også bruksbegrenset til FoU. Mengden av parallelltekster er for liten til statistisk basert maskinoversetting.

Mye av tekstene er merket eller kan enkelt merkes opp med bruk av automatiske taggere (for eksempel Oslo-Bergen-taggeren). Det er imidlertid stor mangel på annotert (lingvistisk oppmerket, f.eks. med hensyn på ordklasse) tekst der merkingen er manuelt kontrollert og korrigert. Dette materialet er nødvendig, bl.a. som grunnlagsdata for å utvikle forbedrete automatiske oppmerkingsverktøy.

En oppdatert oversikt over eksisterende tekstdata er gitt i Tabell B.2 i Vedlegg B. Tabell 4 i kapittel 4 inkluderer en sammenstilling av tilgjengelige tekstressurser fordelt på typer. Når det gjelder enspråklige tekster med basal tilrettelegging, inneholder materialet fra NST om lag 750 millioner ord. Begrensninger på bruk og manglende balanse i materialet gjør at det er anslått å tilsvare 150 mill. ord. I tabellen er det også angitt tall for minimumsinnhold og ønsket innhold i Språkbanken. Merk at behovstallene er per målform, dvs. at for å dekke både bokmål og nynorsk må tallene multipliseres med to. Tall for eksisterende data er totalttall og dekker begge målformer.

2.3 Leksikalske data

I NST-materialet er det også et stort bokmålsleksikon som er godt merket. Leksikonet er basert på avistekster fra seint 1990-tall supplert med Bokmålsordboka fra samme periode, navnelister (ONOMASTICA) samt genererte former laget med en inflektor. Totalt er det ca. 750 000 ordformer i leksikonet. Dette må oppdateres med nyere materiale som må transkriberes og merkes med metadata. Et minimum er at man legger inn en ny versjon av Bokmålsordboka, og Norsk ordbank (nyordsbasen) som UiO arbeider med, vil være et godt supplement. For flere av de identifiserte kildene vil

rettighetene måtte avklares, og det kan være at noen av aktørene kan ha krav på godgjørelse for ressurser som er skapt på grunnlag av andre (avledede data).

NorKompLeks-leksikonet har også en nynorsk ordliste basert på Nynorskordboka som kan danne en grunnstamme for et nynorskleksikon. I tillegg eksisterer nynorskordlister i forlagseie og ressurser utviklet gjennom prosjektet med Norsk Ordbok som kan supplere dette materialet.

Generelt har det kommet til mye leksikalsk materiale etter 2002. Det er særlig universitetene som har laget nye leksikalske baser til forskningsformål. Disse er godt strukturert, følger tilgjengelige standarder og er tilrettelagt for gjenbruk. Innlemming i Språkbanken kan kreve noe innsats med opprenskning og tilpassing til felles format og ev. justering med hensyn til oppmerking, men det bør være et overkommelig arbeid som kan utføres i fagmiljøene.

Det finnes også store mengder fagterminologi som er elektronisk lagret, og som er av interesse for Språkbanken. Det er særlig dikteringssystem som har bruk for denne typen språkressurser. Et første trinn her kan være å kartlegge hva som finnes, og hvilken kvalitet de ulike ressursene har. Det er nødvendig med en faglig oppdatering av eldre materiale samt at alt som tilbys gjennom Språkbanken, må standardiseres. Den delen av arbeidet må gjøres av fagmiljøene, men kan f.eks. koordineres av Språkrådet, noe som kan medvirke til at Språkrådet får et godt utgangspunkt til å ta fatt på de nye oppgavene i terminologi som Språkrådet nå har fått (jf. St.meld. nr. 35 (2007–2008) s. 109. Terminologisamlingene som finnes, er av varierende kvalitet og omfang, fra enkle en- og flerspråklige ordlister til avanserte kunnskapssystem med omfattende informasjon om hver term og frase.

En oppdatert oversikt over eksisterende leksikalske ressurser er gitt i Tabell B.3 i Vedlegg B. Tabell 5 i kapittel 4 inkluderer en sammenstilling av tilgjengelige leksikalske ressurser fordelt på typer.

2.4 Verktøy

Det er utviklet en rekke relevante verktøy, primært ved universitetene. Svært mange av disse verktøyene er produsert i forbindelse med offentlig finansierte forskningsprosjekter, for eksempel via KUNSTI-programmet i Norges forskningsråd. Hoveddelen av disse ressursene vil i utgangspunktet være tilgjengelige for distribusjon gjennom Språkbanken og vil kunne utgjøre viktige, kostnadsbesparende verktøy ved produksjon av nye språkressurser.

Taggere for grammatisk oppmerking av tekst (NoTa-taggeren og Oslo-Bergen-taggeren) er viktige verktøyressurser. Med tilgang på nytt, manuelt kontrollert og korrigert tekstmateriale vil det være ønskelig å forbedre kvaliteten på de automatiske taggerne og dermed redusere behovet for manuell innsats i produksjon av framtidige

databaser. Dataverktøy for produksjon av trebanker (tekstkorpus der setningsstrukturen (dataene) er ordnet i en trestruktur) og verktøy utviklet i forbindelse med oppbyggingen av Norsk aviskorpus vil også være nyttige i produksjonen av nye tekstbaser.

Det er også verdt å nevne at NTNU har laget verktøy som brukes til utvikling av automatisk talegjenkjenning og for høykvalitets talesyntese er finansiert av KUNSTI-programmet og vil være tilgjengelig for innlemmelse i Språkbanken.
En samlet oversikt over eksisterende verktøy er gitt i Tabell B.4 i Vedlegg B.

3 Innhold i Språkbanken – norsk BLARK

Internasjonalt har det utviklet seg en standard for hva som trengs i en språkbank som skal være tjenlig for språkteknologisk forskning og utvikling. Konseptet kalles BLARK (Basic LAnguage Resource Kit), og dette brukes til å identifisere hva vi har og hva vi trenger i årene framover for å gjøre den norske språkbanken formålstjenlig.

3.1 Hva er BLARK?

BLARK (*Basic LAnguage Resource Kit*) er en systematisk framstilling av språkressurser som utgjør en minimal samling, nødvendig for å drive med forskning, utvikling og utdanning i språkteknologi for et gitt språk. Den første BLARK-definisjonen ble gjennomført for nederlandsk språk. ELRA (*European Language Resource Association*), som er den sentrale europeiske distributøren for språkressurser, har adoptert BLARK-konseptet som et universelt grunnlag for å identifisere språkressursbehovet for ulike språk.

Prosedyren med å utarbeide en BLARK starter med å identifisere et sett av viktige språkteknologiske anvendelser. Så foretas en analyse som har som formål å identifisere de verktøy (språkteknologiske moduler) og språkdata som er nødvendige for å realisere hver enkelt av disse anvendelsene. En vurdering av tilgjengeligheten til hver enkelt modul og datatype gir til slutt et bilde av status for språkteknologiske ressurser.

Verktøykassa av språkressurser som en BLARK utgjør, kan altså deles inn i en definisjonsdel og en innholdsdel. Definisjonsdelen er en oversikt over viktige språkteknologiske moduler og hvordan de er avhengige av tilgang på ulike datatyper. I tillegg inneholder den en oversikt over viktige anvendelser og deres avhengighet av de ulike språkteknologiske modulene. Innholdsdelen er en "statusrapport" om i hvilken grad de ulike datatypene og språkteknologiske moduler eksisterer og er tilgjengelige.

En BLARK-beskrivelse vil utgjøre et viktig element i grunnlaget for å utarbeide en prioriteringsliste for en norsk språkbank.

Vi har utarbeidet et første utkast til en norsk BLARK-beskrivelse. Mye av definisjonsdelen av BLARK vil være relativt språkuavhengig, og utgangspunktet for den norske versjonen er primært den nederlandske BLARK-beskrivelsen.

3.2 Beskrivelse av en norsk BLARK

Med utgangspunkt i BLARK-spesifikasjonene for nederlandsk er 13 viktige anvendelser av språkteknologi identifisert. Disse er:

- dataassistert språklæring
- adgangskontroll
- talestyring
- automatisk transkripsjon
- diktering
- tekst-til-tale-syntese
- dialogsystemer
- dokumentproduksjon
- automatisk sammendrag
- informasjonsgjenfinning
- informasjonsaksess
- oversetting (tekst til tekst)
- oversetting (tale til tale)

For hver av disse anvendelsene foretas en analyse av hvilke språkteknologiske moduler som er nødvendige, og hvor viktige de er (f.eks. talegjenkjenning, akustiske modeller, morfologisk analyse osv.). Så foretas en analyse av graden av betydning for de dataressursene som er nødvendige for å realisere de ulike språkteknologiske modulene. Analysen resulterer i et sett av tabeller som er presentert i Vedlegg A. Det må understrekes at de ulike kategoriene i tabellene vil inneholde et vidt spenn av annoteringsdetaljer, datatyper og kvalitet og derfor må betraktes som en makrobeskrivelse.

I innholdsdelen er tilgjengelighet vurdert på en skala fra 1–9, der 9 tilsvarer at ressursen er fullt tilgjengelig i ønsket kvalitet og kvantitet, mens 1 tilsvarer total mangel på tilgjengelige ressurser.

Tallfestingen av tilgjengelighet for hver enkelt ressurs er basert på en skjønnsmessig vurdering av faktorer som mengde og kvalitet på dataene i forhold til ønsket nivå (se kap. 4), tilgjengelighet til data, grad av standardisering, grad av bearbeiding og oppmerking. Karaktersettingen i denne første versjonen må kvalitetssikres. Selv om sentrale fagpersoner i det norske språkteknologimiljøet har bidratt med synspunkter og kommentarer, vil det være behov for en omfattende høringsprosess for å få definert en omforent norsk BLARK. For å kunne benytte det norske BLARK-innholdet i en detaljert planleggings- og prioriteringsprosess er det nødvendig å ha en mer finkornet karaktersetting, der en i tillegg til en totalkarakter har med en vurdering av kvalitet, grad

av oppmerking, begrensninger på bruk osv. Det vil være sterkt ønskelig å gjennomføre en slik prosess i løpet av høsten 2008.

Tabell 1 og 2 viser grad av tilgjengelighet i forhold til ønsket nivå for henholdsvis språkdata og språkmoduler.

Data	Beskrivelse	Tilgjengelighet
Enspråklige leksika	Enspråklige ordlister med ord og bøyningsinformasjon, uttalebeskrivelse	6
Navneleksika	Egennavn, stedsnavn	6
Flerspråklige leksika	Ordlister for oversettelse	4
Tesaruser	Leksikon med rordelasjoner	3
Ontologier, ordnett	Semantiske ordlister og leksika	2
Ikke-annotert tekst	Umerket og maskinannotert tekst	3
Annotert tekst	Manuelt kontrollert maskinannotert tekst	2
Manuskriptlest tale	Planlagt tale, rom- og telefonkvalitet	6
Spontan tale	Naturlig tale, rom- og telefonkvalitet	4
Talte dialoger++	Intervjuer, dialoger, møteopptak osv.	4
Flerspråklig tekst	Parallelle tekster på to eller flere språk	2
Multimedie/-modale korpora	Audiovisuelle opptak for språkteknologi, opptak med flere interaksjonstyper (tale, tekst, datamus osv.)	2

Tabell 1. Tilgjengelighet av dataressurser

Tekstmoduler			Talemoduler		
Modul	**Beskrivelse**	**Tilgjenge-lighet**	**Modul**	**Beskrivelse**	**Tilgjenge-lighet**
Grafem-til-fonemkonvertering	Fra ortografisk skrift til uttale	7	Komplett talegjenkjenner	Verktøy, data, prosedyrer og modeller	6
Token-deteksjon	Gruppering av tekstsymboler i meningsfylte enheter.	3	Akustiske modeller	Statistiske modeller av lydproduksjon	6
Deteksjon av setningsgrenser	Identifisering av hvor setninger starter og slutter	3	Språkmodeller	Statistiske modeller av språkstruktur	3
Navnegjenkjenning	Deteksjon av egennavn	4	Uttaleleksikon	Beskrivelse av uttale	7
Stavekorreksjon	Forslag til riktig staving	6	Robust talegjenkjenning	Talegjenkjennning i støy	4
Lemmatisering	Grunnform-identifisering	4	Talegjenkjenning for dialekter og innvandrernorsk	Talegjenkjenning av naturlig dagligtale	1
Morfologisk analyse	Ordanalyse	8	Taleradapsjon	Brukertilpasnlng av statistiske modeller	5
Morfologisk syntese	Generering av fullformsord	6	Leksikalsk adapsjon	Brukertilpasning av uttale	4
Ord-disambiguering	Avklaring av flertydighet	6	Prososdi-gjenkjenning	Intonasjon, trykk som påvirker tolkning av tale.	4
Parsere, grammatikker	Setningsanalyse og språkbeskrivelser som brukes i automatisk setninganalyse	7	Komplett talesyntese	Verktøy, data, prosedyrer og programvare for tekst-til-tale.	6
Grunn parsing	Omtrentlig setningsanalyse	2	Difonsyntese	2.generasjons talesyntese.	8
Konstituent-gjenkjenning	Hiearkisk grammatisk analyse	4	Skjøtesyntese	Høykvalitets talesyntese	6
Semantisk analyse	Maningsanalyse	2	Prosodiprediksjon for TTS	Intonasjons-modul for talesyntese	4
Referentanalyse	Forankring av f.eks. pronomen til refererende uttrykk, f.eks. egennavn	3	Automatisk fonetisk transkripsjon	Oppmerking av taledatabaser	6
Pragmatisk analyse	Kontekstuell meningstokning	1	Automatisk fonetisk segmentering	Detaljert oppmerking av taledatabaser	6
Tekstgenerering	Produksjon av tekst fra konsepter	3	Fonetiske likhetsmål	Tallfesting av ulikhet mellom språklyder	5
Språkgjenkjenning (tekst)	Hvilket språk er teksten på?	5	Talergjenkjenning	Bestemmelse av talerens identitet	3
Språkavhengig oversetting	Maskinoversettelse mellom bestemte språkpar	3	Talersporing	Sppring av talerskifte i opptak med flere talere	2
Sidestilling av parallelltekst	Automatisk lenking av ord mellom tekster	6	Språkidentifkasjon	Hvilket språk snakkes?	2
POS-tagger	Program som bestemmer ordklassen til hvert ord i en tekst	7	Dialektidentifikasjon	Hvilken dialekt snakkes?	2
Term-ekstraksjon	Verktøy for korpusbasert terminologi-oppbygging	4	Konfidensmål	Gradering av sikkerhet av gjenkjennings-resultat	4
			Ytringsverifikasjon	Kontroll av at et dialogsystem har forstått brukeren	3
			Emosjonsidentifikasjon	Tolkning av talerens sinnstilstand.	1
			Taledeteksjon	Skille tale fra bakgrunnsstøy, musikk osv.	4

Tabell 2. Tilgjengelighet av tekst- og talemoduler

4 Analyse – konsekvenser av BLARK

I kapittel 2 ble det presentert en oversikt over eksisterende, tilgjengelige språkressurser. Her ble det også gitt anslagstall for det nødvendige minimumsinnholdet av dataressurser i Språkbanken. BLARK-analysen gir en oversikt over status etter dagens behov. Disse oversiktene danner grunnlaget for planlegging av oppbyggingen av Språkbanken.

Det er nødvendig å påpeke at etablering av Språkbanken vil måtte følges opp av en kontinuerlig vurdering av nye behov. Språket er i stadig utvikling, og det er viktig at ressurssamlingene suppleres med data som avspeiler dette. I tillegg gjør utvikling av såvel samfunnet som teknologien at nye bruksområder for språkteknologi vil gjøre det nødvendig med supplerende data. Eksempler på dette er bruken av SMS-språk og den mer muntlige tekstformen som benyttes i e-postmeldinger, og at antallet innbyggere som har norsk som andrespråk, vil gjøre det nødvendig å ha språkteknologiske verktøy som håndterer "innvandrernorsk" på lik linje med norske dialekter.

I vurderingen av ressursene kan det være nyttig å ta hensyn til noen faktorer som spesifiserer begrensinger for tilgjengelighet. *Svart-boks data* er data eller verktøy der en ikke kan inspisere innholdet, eller gjøre endringer, *glass-boks data* er data eller verktøy der innholdet kan inspiseres, men ikke endres. *Åpne ressurser* har ingen begrensninger på bruk.

4.1 Taledata

Taledata er kjernen i all teknologi som omhandler gjenkjenning av tale og produksjon av tale. Talegjenkjenning krever opptak av mange talere i ulike aldersgrupper og med ulike dialekter, og opptakene bør være knyttet til realistiske brukssituasjoner. Opptakene må som et minimum være ortografisk transkribert. For syntetisk tale kreves opptak av en enkelt taler per syntetisk stemme. Opptakene må være av høy kvalitet og ha best mulig dekning av naturlig forekommende lydkombinasjoner og intonasjoner. Talemateriale for syntetisk tale må være detaljert oppmerket.

Tilgangen på eksisterende taledata er kraftig forbedret siden 2002, i hovedsak på grunn av oppkjøpet av konkursboet etter NST. Dette materialet inneholder manuskriptlest bokmålstekst, både med romkvalitet og telefonkvalitet, og dekker mye av behovet for mansukriptlest tale. Det er imidlertid fortsatt behov for å supplere med manuskriptlest nynorsktekst og med tale fra barn/ungdommer og eldre. For spesielle formål som medisinsk diktering er det materiale både fra NST-boet og data innsamlet av Max Manus som kan brukes. Rettigheter og kompensasjon for bruk av slikt materiale er imidlertid ikke avklart.

Når det gjelder spontantale, utgjør NoTa-korpuset et godt startpunkt. Dette materialet dekker imidlertid bare Oslo-dialekt, og er dialoger. Også Rundkast-korpuset med

kringkastede nyhetssendinger inneholder en del spontantale. Det er imidlertid stor mangel på spontane monologer (for eksempel for diktering) og databaser med god dekning av norske dialekter.

Det er heller ikke samlet inn tilstrekkelige mengder med realistiske taledata som kan benyttes til opptrening av talte dialogsystemer. Dette vil i hovedsak involvere telefontale. Telenor har transkriberte opptak av om lag 100 000 samtaler fra 05000-tjenesten. Det er imidlertid uvisst om dette er data som kan gjøres tilgjengelig for Språkbanken.

Det er samlet inn mye data for høykvalitets talesyntese gjennom Fonema-prosjektet ved NTNU. Taledata for høykvalitets talesyntese er også tilgjengelig fra NST-boet. Prosedyrene for manuskriptgenerering og opptak fra Fonema-prosjektet vil kunne gjøre det mulig med en effektiv supplering av taledata for talesyntese.

Behovsvurderingene for taledata er vist i Tabell 3 og er i stor grad samsvarende med de tallene som ble utarbeidet i 2002. Minimumstallet for manuskriptlest telefontale er økt fra 120 til 500 timer, men de 500 timene tilsvarer det tilgjengelige talematerialet fra NST-boet som bør innlemmes i Språkbanken, og som ikke vil innebære noen vesentlig kostnadsøkning. Omfanget av spontantale er noe redusert, først og framst av hensyn til kostnadene.

				Minimum	Ønsket
Type	Talestil	Formål	Timer, tilgjengelig	Timer, minumum	Timer, ønsket
Romkvalitet	Spontan	Diktering, dialoger	100	450	1000
Romkvalitet	Manuskript	Diktering, modeller	540	500	1000
Telefon	Manuskript	Modeller	500	500	700
Mobiltlf	Manuskript	Modeller	200	120	240
Tlf i bil	Manuskript	Diverse	0	120	240
Telefon	Spontan	Dialoger	0	100	200
Telefon	Spontan	Diktering	0	100	200
Romkvalitet	Manuskript	Difondatabase	0	2	4
Romkvalitet	Manuskript	Prosodi / Lydbibliotek	1	20	40
Kringkasting	Variert	Emnesøk	70	70	200
Telefon	Manuskript	Emnesøk i multimedia-arkiver	0	20	40
Audio	Spontan	Emnesøk	20	100	200
Romkvalitet	Spontan	Multimodale grensesnitt	30	30	100
Romkvalitet	Spontan	Modeller, flerspråklige applikasjoner	0	0	100
Høy romkvalitet	Manuskript	Konkatenativ talesyntese	20	40	80
SUM			1 481	2 172	4 344

Tabell 3. Taledata, behov

4.2 Tekstdata

Omfattende tekstdatabaser er en primærkilde for utvikling av leksikalske ressurser og for statistiske språkmodeller for talegjenkjenning. Slike tekstbaser må ha god dekning av ulike typer tekster (sakprosa, småtrykk, aviser, skjønnlitteratur osv.) og være automatisk ordklassemerket. Flerspråklige parallellkorpus er en nødvendighet for maskinoversetting. For statistisk basert maskinoversetting trengs også her omfattende databaser. Tekstbaser med grundig merking (trebanker, utvidet merking som er manuelt kontrollert) er nødvendige for syntaktisk og semantisk analyse og for utvikling av bedre verktøy for automatisk merking.

For tekstdata er det en markert ubalanse mellom bokmål og nynorsk. For basalt oppmerket bokmålstekst har vi tekstbasen fra NST-boet og Norsk aviskorpus som til sammen utgjør om lag 1,4 milliarder ord løpende tekst. Dette tekstmaterialet er ikke tilstrekkelig balansert, det er en stor overvekt av avistekst. Det er også uklarheter knyttet til opphavsrett og bruksrett for NST-dataene. For Norsk aviskorpus må eventuell innlemmelse i Språkbanken avklares med avisene. Sannsynligvis kan disse to samlingene uansett benyttes som en svartboksressurs, for eksempel til å generere statistiske språkmodeller (N-gram) som kan distribueres gjennom Språkbanken. Kvaliteten til slike modeller forbedres med økt mengde av treningsdata. Det kan her nevnes at Google har tilgjengeliggjort N-gram for engelsk basert på et datasett på vel en milliard ord i løpende tekst.

Basal ordklassemerking av store databaser kan i stor grad gjøres maskinelt. Imidlertid er det behov for databaser som har manuell kontroll og retting av den maskinelle oppmerkingen. Slike databaser vil være nødvendige for å forbedre presisjonsnivået til taggerne og for å muliggjøre selvlærende systemer.

En del tospråklige tekster er produsert via KUNSTI-programmets maskinoversettingsprosjekt, LOGON. Disse tekstene er tilgjengelige for Språkbanken, men er imidlertid innen et svært avgrenset tematisk domene. I tillegg eksisterer Oslo Multilingual Corpus på om lag 15,5 millioner ord som imidlertid har opphavsrettslige begrensninger som kan utelukke samlingen fra Språkbanken. Mengden av parallelltekster er uansett vesentlig for liten til bruk i statistisk basert maskinoversetting. Verktøy som for eksempel Aksis sin Text Corpus Aligner for parallellstilling av parallelle tekster er en nødvendighet for produksjon av større, flerspråklige parallelltekstbaser.

Tilfanget av tekstbaser med detaljert oppmerking er relativt beskjedent. Det er imidlertid utviklet en del verktøy, for eksempel for produksjon av trebanker, som vil kunne effektivisere produksjonen.

Behovsvurderingen for tekstdata er vist i Tabell 4. Merk at behovstallene er per målform, dvs. at for å dekke både bokmål og nynorsk må tallene multipliseres med 2. Tall for eksisterende data er totaltall og dekker begge målformer. Behovene samsvarer i

stor grad med vurderingene som ble gjort i 2002, med unntak av behovet for basalt oppmerkede tekstdata, der minimumsomfanget er økt fra 50 millioner til 250 millioner ord i løpende tekst per målform. Vurderingene fra 2002 betraktes som altfor lave i forhold til de krav dagens teknologi setter. En kan merke seg at mengden tilgjengelige tekstdata er angitt som betydelig mindre enn størrelsen av NST-dataene og Norsk aviskorpus. Dette skyldes at vi har gjort en skjønnsmessig avkorting på grunn av usikkerhet om tilgang og eierrettigheter og for å korrigere for at dette materialet er dårlig balansert med hensyn til teksttyper. Minimumsmengden av tospråklige tekster er også for liten til statistisk basert maskinoversetting og må betraktes som et absolutt minimum.

			Minimum	Ønsket
Teksttyper	Bearbeiding	Tilgjengelig (størrelse)**	Størrelse*	Størrelse*
Tospråklige tekster (norsk - engelsk)	Basal tilrettelegging	175 000	2 500 000	5 000 000
Tospråklige tekster (engelsk - norsk)	Basal tilrettelegging	175 000	2 500 000	5 000 000
Tospråklige tekster (engelsk - norsk og norsk - engelsk)	Grundig tilrettelegging	0	500 000	1 000 000
Sakprosa, småtrykk, upublisert materiale	Basal tilrettelegging	150 000 000	250 000 000	1 000 000 000
Aviser og media, skjønnlitteratur	Basal tilrettelegging	150 000 000	250 000 000	1 000 000 000
Sakprosa, småtrykk, upublisert materiale	Utvidet tekstkoding, manuelt kontrollert ordklassemerking	0	500 000	1 000 000
Aviser og media, skjønnlitteratur	Utvidet tekstkoding, manuelt kontrollert ordklassemerking	0	500 000	1 000 000
Aviser	Etablering av trebank	0	200 000	1 000 000
Anonymiserte journaler	Treningsdata for medisinsk diktering	0	0	200 000
SUM		**300 350 000**	**506 700 000**	**2 014 200 000**

Tabell 4. Tekstdata, behov. (*Pr målform **Totalt, bokmål og nynorsk)

4.3 Leksikalske data

Denne delen av språkbanken vil inneholde leksikon og tesauruser. Med leksikon mener en her elektroniske ordlister med informasjon om ordforrådet i et språk på ulike språkvitenskapelige nivå. Tesauruser er leksikon med semantiske og assosiative relasjoner mellom ord, inklusive emnetesauruser for fagterminologi f.eks. fra medisin.

Det eksisterer en relativt stor mengde med leksikalske grunnordlister. Ordlistene fra NST-boet inneholder om lag 750 000 fullformer for bokmål. Hoveddelen av dette materialet er egenutviklet, men det er også benyttet innkjøpte ressurser til å supplere egne data. Dette gjør at det er et behov for avklaring av rettighetsspørsmål før dette leksikonet kan innlemmes i Språkbanken. NorKompLeks-leksikonet er fullformsleksika for både bokmål og nynorsk og kan trolig inngå i Språkbanken. Telenor har en eierandel i NorKompLeks og eventuelt vederlag må avtales. Telenor har også rettighetene til navneleksikonet ONOMASTICA, som er svært aktuelt for innlemming i Språkbanken.

Nynorskdelen av NorKompLeks-leksikonet er basert på Nynorskordboka og kan danne en grunnstamme i et nynorskleksikon, og det er ønskelig å innlemme materiale fra Norsk Ordbok i Språkbanken etter hvert som det blir tilgjengelig. I tillegg finnes det nynorskordlister i forlagseie som kan supplere dette materialet.

Språkmeldinga tar til orde for at norsk terminologi må utvikles raskt og effektivt og "stillast til rådvelde for alle aktuelle brukarar gjennom digitale terminologibasar som er allment tilgjengelege over Internett". Så vel eksisterende terminologilister som nyutviklede terminologiressurser bør tilrettelegges for språkteknologisk bruk og distribueres gjennom Språkbanken.

Det er for øvrig en stor mengde med relevante, leksikalske data som eies av ulike forlag. Dette må undersøkes nærmere med de enkelte rettighetshavere. Kunnskapsforlaget, som eier en stor mengde ordbokressurser, er i prinsippet velvillig til å bidra til Språkbanken, men her må betingelsene avklares nærmere. Tospråklige ordbøker er spesielt viktige for maskinoversetting.

Behovsvurderingene for leksikalske data er vist i Tabell 5. Disse tallene samsvarer med vurderingene fra 2002. Utvikling av stavevarianter og uttalebeskrivelser samt innlemming av eksisterende ordlister er ikke tallfestet i form av antall ord, men vil ha utgifter til bearbeiding og tilrettelegging samt utgifter til frikjøp av rettigheter (se kapittel 7).

	Eksisterende	**Minimumsbehov**	**Ønsket behov**
Aktivitetstype	Antall ord (fullformer)	Antall ord (fullformer)	Antall ord (fullformer)
Kjøp og kvalitetskontroll ordlistedata, bokmål	750 000	500 000	1 000 000
Kjøp og kvalittskontroll ordlistedata, nynorsk	300 000	500 000	1 000 000
Innlemming av ordlister fra ulike kilder	0		
Utvikling av stavevarianter/basis dialektvarianter	0		
Utvikling av uttalebeskrivelse for navn, fremmedord og nyord	0		
Uttalebeskrivelser for dialektregioner	0		
Tospråklige parallellordlister	75 000	75 000	100 000
Ordnett (norsk Wordnett)	1 000	50 000	100 000
Begrepsbeskrivelser - SIMPLE	10 000	50 000	100 000
SUM	**1 136 000**	**1 175 000**	**2 300 000**

Tabell 5. Leksikalske data – behov

4.4 Verktøy

Rapporten fra 2002 hadde ikke med en eksplisitt oversikt over verktøy som ressurs i en språkbank, men slo fast at "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". De fleste språkteknologiske fagmiljøene her i landet og internasjonalt har siden den tid utviklet ulike verktøy som er aktuelle å tilby sammen med språkressurser fra Språkbanken. Det kan være verktøy som renser tekst for skrivefeil, det kan være verktøy for merking av tekst, det kan være verktøy for konvertering til/fra ulike format osv. De verktøyene som er aktuelle i denne sammenhengen, vil være verktøy som må til for at brukerne skal kunne nyttiggjøre seg språkressursene til sine formål. De samme språkressursene kan bli anvendelige for ulike brukere såframt de kan få tilgang til verktøyene også.

Internasjonalt er det et utstrakt samarbeid når det gjelder utvikling av språkuavhengige verktøy, og det norske språkteknologimiljøet har kompetanse som gjør at det i stor grad deltar i internasjonalt samarbeid på dette området også.

En del verktøy er kommentert under avsnittene om tale, tekst og leksikalske data. I Tabell B.4 i vedlegg B er det gitt en oversikt over eksisterende verktøy som er tilgjengelige fra ulike institusjoner. Fra denne tabellen og fra BLARK-analysen vil det framgå at det er bra tilgjengelighet for enkelte verktøy, mens andre verktøy er ikke-eksisterende eller har svært dårlig tilgjengelighet. Verktøyutvikling som er nødvendig for produksjon, oppmerking, tilgang og analyse av språkdata, er innkalkulert i kostnadsestimatene for de ulike typer språkdata. Andre verktøy forutsettes utviklet gjennom språkteknologiske forskningsprosjekter. Det er essensielt at etablering av en norsk språkbank følges opp med en parallell satsing på språkteknologisk forskning, og det vil være naturlig at en del av verktøyutviklingen inngår i denne satsingen.

4.5 Validering og kvalitetskontroll

Fagmiljøene kjenner til og benytter vanligvis de internasjonale standardene for annotering og tilrettelegging som er i bruk internasjonalt. Alt materiale som tilbys gjennom Språkbanken, må være kvalitetssikret og validert av en ekstern, nøytral part. Den europeiske distributøren av språkressurser i Paris, ELDA, vil ut fra sin posisjon i det internasjonale arbeidet med distribusjon av språkressurser og sitt arbeid med standardisering på området være en premissleverandør og normgiver.

Taleressursene i NST-materialet ble i sin tid kvalitetssikret internt, men de trenger en uavhengig gjennomgang. Dette er en oppgave som kan iverksettes fra dag en i Språkbankens regi. Når det gjelder taleressurser generelt, er SPEX (akronym for Speech Processing Expertise Centre) i Nederland en aktuell kandidat. Denne gruppen validerer og kvalitetssikrer alt talemateriale som tilbys gjennom ELDA.

4.6 Forberedende arbeid høsten 2008

Arbeidsgruppen har fått inn en grov oversikt over ressurser som har kommet til hos ulike aktører etter 2002. Vi har tatt utgangspunkt i oversikten over eksisterende materiale fra 2002 og supplert den. Oversiktene må gås gjennom på nytt og suppleres med en mer detaljert gjennomgang og nærmere vurdering av det som ble rapportert inn i løpet av sommeren 2008. Oversiktene i 2002 var, som nå, basert på egenrapportering fra leverandørene. Det har ikke vært tid til å vurdere materialet i detalj med hensyn til relevans, kvalitet eller hva det vil kreve av arbeidsinnsats for å gjøre det tilgjengelig for språkteknologisk bruk.

Høsten 2008 bør brukes til en grundigere kartlegging av identifiserte ressurser, hva som finnes hvor, hvilke opphavsrettlige problemstillinger som er knyttet til dem, hvilken kvalitet de har, og eventuelt et overslag over hvor mye det vil koste å gjøre dem tilgjengelige for brukerne av Språkbanken. Det bør også gjennomføres et grundigere arbeid med en norsk BLARK-beskrivelse som vil kunne gi et bedre grunnlag for prioritering av nyinnsamling. Alt eksisterende materiale må valideres og kvalitetssikres for å få en oversikt over om det holder internasjonale mål og følger internasjonale standarder eller "beste praksis". En slik gjennomgang vil resultere i en mer realistisk beregning av hvor mye midler som må settes av til bearbeiding av sentrale ressurser for å gjøre dem tilgjengelige for gjenbruk. Vurderingen må holdes opp mot prioriteringene som kommer fra fagmiljøene, og veies opp mot kostnadene ved nyinnsamling av tilsvarende materiale.

Oppgavene med en detaljert gjennomgang av hva som finnes hvor, og hvilken kvalitet de ulike ressursene har, ligger utenfor arbeidsgruppens oppdrag. Arbeidsgruppen vil likevel anbefale at arbeidet gjøres høsten 2008 for ikke å tape tid og for å forberede oppstarten av Språkbanken på best mulig vis. Arbeidet vil kreve midler til reiser og arrangering av møter for at fagfolk skal kunne levere et mer konkret og detaljert underlag for hvor Språkbanken bør starte arbeidet med innsamling av språkressurser. Arbeidsgruppen anslår at kostnadene vil beløpe seg til om lag 150 000 kr.

Det er en fordel om Kultur- og kirkedepartementet forbereder oppnevning av styret for Språkbanken slik at det kan være på plass så snart Stortinget har gjort de nødvendige vedtak. Så snart styret er på plass, kan man begynne prosessene med tilsetting av personalet, noe som vil gi mulighet for å starte det praktiske arbeidet tidlig i 2009.

4.7 Prioriteringer i 2009 og videre arbeid i 2010–2014

Ved den foreslåtte prioriteringen er det lagt til grunn at det er viktig å få etablert Språkbanken med et innhold som gjør at brukermiljøene tidligst mulig kan få tilgang til viktige språkressurser. Dette innebærer at prioritetsoppgaven i startfasen blir å tilgjengeliggjøre eksisterende språkdata. For arbeid med nyinnsamling må det tas hensyn til at fagmiljøet i Norge er av begrenset størrelse, og at det ikke er ønskelig å

gjøre en midlertidig oppbygging av personellressurser som må nedskaleres når Språkbanken er etablert. Derfor må arbeidet med nyinnsamling balanseres mellom ulike datatyper, slik at de enkelte fagmiljøene ikke blir overbelastet med innsamling og bearbeiding av språkressurser.

I 2009 vil primæroppgavene være å etablere Språkbanken, innlemme de viktigste eksisterende ressursene og starte nyinnsamlingsarbeidet. Etablering av Språkbanken vil inkludere opprettelse av selve organisasjonen, styreutnevnelse, ansettelse av personale, utarbeiding av standard kontraktsformularer osv.

Talematerialet fra NST er den største eksisterende språkressursen, og kjøp av bruksrett, kvalitetssikring og tilretteleggig av dataene for distribusjon vil være den første store oppgaven. Resten av NST-materialet bør også tilgjengeliggjøres, men det må først foretas en avklaring av rettighetsforholdene.

Andre eksisterende ressurser med avklarte rettighetsforhold må videre innlemmes i Språkbanken etter nødvendig tilrettelegging og kvalitetskontroll. Det meste av data og verktøy som er etablert i tilknytning til KUNSTI-programmet og andre forskningsprosjekt finansiert av Norges forskningsråd vil inngå i denne kategorien.

For viktige ressurser som har begrensning på bruksrett, må det forhandles med rettighetshaverne om distribusjon via Språkbanken. Dette gjelder f.eks. Norsk aviskorpus og talekorpus og leksika som eies av Telenor.

For nyinnsamling bør innsamling av manuelt kontrollerte, annoterte tekstkorpus og taledata med spontantale startes opp i 2009.

Det kartleggingsarbeidet som arbeidsgruppen foreslår gjennomført høsten 2008, vil danne grunnlaget for den detaljerte planen for perioden 2010–2014 som må utarbeides i 2009.

I 2010 vil nyinnsamlingsarbeidet som ble startet i 2009, videreføres. De mest aktuelle nye innsamlingsprosjekter som kan startes i 2010, er for tospråklige tekstkorpora for maskinoversetting, basalt annotert tekst utenfor avisdomenet og etablering av norske trebanker. Det bør også forhandles med rettighetshavere om innlemmelse av ordbokdata i Språkbanken.

For perioden 2011–2014 er prioriteringsrekkefølgen mer uklar og den vil bestemmes av planleggingsarbeidet som forutsettes gjennomført i 2009, av kapasitet i fagmiljøene og eventuelle overordnete prioriteringer. Det må samles inn taledata med dialekter og informanter med norsk som andrespråk, grundig merkede tospråklige tekstsamlinger for maskinoversetting, uttalebeskrivelser for dialekter og nyord for å nevne de viktigste elementene.

5 Organisering

St.meld. nr. 35 slår fast at "Det er ingen andre enn Noreg som vil utvikla språkteknologi på norsk. [...] Både i europeisk og i nordisk samanheng tek ein sikte på å oppretta ein forskningsinfrastruktur og eit samarbeid som integrerer tilgjengelege språkressursar. Oppbygging og kvalitetssikring av nasjonale ressursar vil likevel vera ei nasjonal oppgåve." (s. 135).

På grunnlag av St.meld. nr. 35 er det klart at Språkbanken er et statlig tiltak som skal sørge for norskspråklig infrastruktur til forskningsformål og næringsutvikling. Dette er et sektorovergripende kulturpolitisk tiltak som i utgangspunktet ikke innebærer utøvelse av forvaltningsmyndighet.

5.1 Organisasjonsform

I forbindelse med Språkrådets utredning i 2002 ble det innhentet en juridisk betenkning fra Simonsen Føyen Advokatfirma DA: "Betenkning over juridiske problemstillinger knyttet til samling og tilgjengeliggjøring av norske språkteknologiressurser". Den juridiske betenkningen drøftet mulige organisasjonsformer for en språkbank og den konkluderte med tre alternativer: Stiftelse, statsaksjeselskap eller aksjeselskap. Arbeidsgruppen er kommet til samme konklusjon, men legger til grunn at statsaksjeselskap ikke er aktuelt da St.meld. nr. 35 "Mål og meining" slår fast at Språkbanken skal opprettes per 1.1.2009.

Staten står i utgangspunktet fritt med hensyn til organisering av Språkbanken. Språkbanken er ikke tiltenkt forvaltningsoppgaver, og det er derfor hensiktsmessig at den organiseres som et eget rettssubjekt.

Det er også en fordel at Språkbanken er et selvstendig rettssubjekt av hensyn til Språkbankens virksomhet, med inngåelse av avtaler om kjøp og salg av bruksrett til språkressurser og lignende.

Språkbanken bør videre kunne handle relativt raskt og ha fullmakter til å sette ut oppdrag, både i etableringsfasen og når man går over i en driftsfase. Det vil også være en fordel med hensyn til regnskapsførsel at Språkbanken er en selvstendig juridisk enhet.

Både stiftelser og aksjeselskaper er lovregulert med gode modeller og verktøy for styring og kontroll. Kultur- og kirkedepartementet bør være eneaksjonær eller majoritetseier i et eventuelt aksjeselskap, av hensyn til det overordnede språkpolitiske ansvar som er lagt til departementet. Språkbanken bør ikke ha private eiere, jf. Pkt. 6.2. Ved valg av stiftelse som organisasjonsform bør Kultur- og kirkedepartementet være stifter. Begge organisasjonsformene kan ivareta kravene som bør stilles til organisering av en virksomhet som Språkbanken.

Styret for Språkbanken bør uansett organisasjonsform være bredt sammensatt, med representanter for forvaltning, forskningsmiljø og næringsliv fordi disse vil være de sentrale brukerne av ressursene. Blant styremedlemmene bør det være personer med god oversikt over og solid kompetanse om behovene i språkteknologisk industri og FoU, nasjonalt og internasjonalt.

For at Språkbanken skal kunne opptre selvstendig og inngi tillit i markedet når det gjelder å motta, forvalte og distribuere språkdata, bør Språkbanken være løsrevet fra institusjons- eller bedriftsmessige særinteresser.

Det er viktig at Språkbanken får en organisasjon som er levedyktig og gis en rimelig forutsigbar økonomi. Språkbanken må kunne legge langsiktige planer og inngå langsiktige avtaler om innsamling og foredling av språkressurser og verktøy ut fra de behovene og ønskene FoU-miljøene og næringslivet har. Det er også viktig at organisasjonen er dynamisk og har evne til å prioritere og omprioritere basert på endringer i behovene og/eller nye muligheter på området.

5.2 Forskjell mellom aksjeselskap og stiftelse

I aksjeselskaper er generalforsamlingen det høyeste organ, og det er aksjeeierne som gjennom generalforsamlingen har den bestemmende innflytelse. Styret, som velges av generalforsamlingen, forestår forvaltningen av selskapet. Aksjeloven har detaljerte regler om organisering og drift av aksjeselskaper, som i stor utstrekning kan fravikes ved bestemmelser i selskapets vedtekter.

Aksjeeierne kan til enhver tid ved beslutning på generalforsamling med kvalifisert flertall endre selskapets vedtekter og kan også beslutte oppløsning av selskapet.

Forvaltningen av stiftelser hører under styret, som er stiftelsens øverste organ. Stiftelsestilsynet fører tilsyn og kontroll med at forvaltningen av stiftelsene skjer i samsvar med stiftelsens vedtekter og stiftelsesloven.

Omdanning av stiftelser kan bare skje i særlige tilfelle, f.eks. hvis stiftelsens formål ikke kan tilgodeses på grunn av utilstrekkelig kapital i stiftelsen eller at omstendigheter har medført at formålet er blitt åpenbart unyttig, uheldig eller ufornuftig. Omdanning kan foretas av Stiftelsestilsynet. I vedtektene kan det bestemmes at andre (bortsett fra oppretteren av stiftelsen) kan gis myndighet til å omdanne stiftelsen, men vedtaket om omdanning skal da godkjennes av Stiftelsestilsynet.

5.3 Forutsetninger for valg av organisasjonsform

Språkbanken vil ha et allmennyttig formål, og virksomheten forutsettes finansiert hovedsaklig gjennom offentlige bevilgninger og delvis gjennom inntekter fra virksomheten (salg av brukerlisenser o.l.).

Ved valg av organisasjonsform må man ta stilling til om man ønsker en organisasjon der eier(ne) tar aktivt del i driften gjennom beslutninger på generalforsamlingen, eller om man ønsker en organisasjon med en selvstendig stilling, løsrevet både fra den/dem som opprettet organisasjonen og fra den/dem som bidrar med kapital eller andre tilskudd (som f.eks. språkressurser i dette tilfellet).

Stiftelsesformen er velkjent og har vært mye benyttet av offentlige organer for organisering av virksomheter som ikke har et økonomisk formål, f.eks. stiftelser som opprettes for å organisere forskningsinstitutter, museer, omsorgsinstitusjoner osv.

Arbeidsgruppen har imidlertid konkludert med at Språkbanken bør etableres som et aksjeselskap. Dette fordi Språkbanken etableres for ivaretagelse av språkpolitiske formål, og aksjeselskapsformen gir staten mer direkte kontroll med virksomheten enn stiftelsesformen.

5.4 Styringsstruktur og lokalisering

Språkbankens styre bør uavhengig av organisasjonsform bestå av fem til sju personer, med representanter for relevante myndigheter, forskning og næringsliv. Språkbanken bør forvaltningsmessig ligge under Kultur- og kirkedepartementet, og departementet bør utnevne styret, gjerne etter forslag fra de relevante miljøene. Daglig leder bør være styrets sekretær. Det bør vurderes om styret skal ha internasjonal representasjon. Internasjonal kontakt og samarbeid er helt nødvendig for at Språkbanken skal kunne oppfylle sitt formål, ikke minst i etableringsfasen.

For å sikre at behovene for innsamling og tilrettelegging av språkressurser gjenspeiles i Språkbankens prioriteringer, bør man opprette et brukerråd og ha et brukermøte en gang i året der et større antall institusjoner og organisasjoner kan komme med innspill, og der prioriteringene kan diskuteres i et større forum.

Språkbanken vil bli en liten organisasjon med kun tre ansatte, og den bør ikke isoleres faglig. Det betyr at den må plasseres i den ene eller den andre typen fagmiljø, og de aktuelle fagmiljøene er språkteknologiske eller språkfaglige.

Legges Språkbanken til et av de aktuelle språkteknologiske fagmiljøene, kan den bli oppfattet som altfor nært knyttet til ett fagmiljø og dermed ikke få tilstrekkelig tillit hos andre. Samlokalisering med en nøytral tredjepart som også kan tilby et aktuelt fagmiljø, kan være en løsning som ivaretar Språkbankens autonomi og integritet i forhold til fagmiljøene, og det kan være en fordel overfor næringslivet. Språkbanken skal være nøytral og like tilgjengelig for alle. Arbeidsgruppen mener at en samlokalisering med

Språkrådet kan være en god løsning, i samsvar med den utvidede rollen Språkrådet er tiltenkt. Det kan likevel tenkes også andre løsninger for lokalisering, ut fra arbeidsgruppens prioriteringer.

5.5 Oppsummering og anbefaling

Etter arbeidsgruppens oppfatning bør Språkbanken opprettes som en selvstendig juridisk enhet under Kultur- og kirkedepartementet, ut fra de overordete språkpolitiske mål etablering av Språkbanken er en del av, og av hensyn til de oppgavene Språkbanken skal utføre. Språkbanken bør være uavhengig overfor ulike sektorinteresser for å sikre legitimitet og få det handlingsrommet Språkbanken vil ha behov for.

Arbeidsgruppen har sett nærmere på to mulige organisasjonsformer: Stiftelse og aksjeselskap. Begge tilfredsstiller behovet Språkbanken har for handlingsrom, og begge vil gjennom regulering i vedtektene kunne ivareta de funksjoner organisasjonen må ha.

Arbeidsgruppen vurderer imidlertid aksjeselskap som best egnet for løpende ivaretakelse av språkpolitiske formål. Eieren vil ha større mulighet for styring og påvirkning i et aksjeselskap enn i en stiftelse der styret etter loven er mer autonomt.

6 Juridiske problemstillinger

Valg av organisasjonsform er drøftet i overstående kapittel, og i dette kapitlet gjennomgås andre aktuelle problemstillinger som må avklares eller hensyntas i forbindelse med etablering og drift av Språkbanken.

6.1 Opphavsrettigheter til språkressurser

Arbeidsgruppen tiltrer den juridiske betenkningen fra 2002 hva gjelder hvilke former for rettigheter som kan foreligge til de språkressursene som skal samles i Språkbanken. Den juridiske utredningen peker på de relevante opphavsrettslige aspekter ved materiale som skal inkluderes i en framtidig språkbank. Både opphavsrett og de såkalte naborettighetene vern for kringkastingssendinger, vern av utøvende kunstneres framføringer, databasevern og katalogvern kan være aktuelle.

Når det gjelder ovennevnte betenknings problematisering av Språkbankens rolle som "formidler" eller "eier" av språkressurser, legger arbeidsgruppen til grunn at det vil være mest hensiktsmessig at Språkbanken inntar en form for mellomposisjon, ved at Språkbanken innhenter ikke-eksklusive rettigheter til bruk, bearbeiding og videreformidling av språkressursene. Ved å innta en slik mellomposisjon vil de ulike eierne av språkressurser fortsatt kunne disponere sitt materiale, samtidig som Språkbanken får de

rettigheter som er påkrevet for Språkbankens formål. En slik mellomposisjon vil antagelig også lette innhentingen av språkressurser. Språkbankens bruksrett kan bestå ved siden av og som en begrensning i den opprinnelige eiers rettigheter. Språkbanken behøver da ikke frikjøpe alle rettigheter til allerede eksisterende materiale, og den opprinnelige opphavsmann kan fortsette sin utnyttelse som før, ved siden av Språkbankens bruk.

I forbindelse med innhenting av materiale til Språkbanken må det vurderes hva slags rettigheter materialet er beheftet med, men det er ingen ting til hinder for å inkludere alle former for rettigheter i én bruksrettsavtale. Bruksretten bør etter det arbeidsgruppen kan se, være lik i alle tilfeller der Språkbanken selv ikke er opphavsmann.

Språkbankens bruksrett må reguleres i avtaler mellom Språkbanken og de opprinnelige rettighetshavere. Slike avtaler bør inngås med utgangspunkt i én standardavtale som kan benyttes for all innhenting av data til språkbankformål. Slik standardavtale om Språkbankens bruksrett til de språkressurser som Språkbanken skal bestå av, bør utarbeides før 1.1.2009. Som det framgår i den juridiske utredningen, må bruksretten omfatte rett til eksemplarframstilling, kopiering, bearbeiding, hel eller delvis viderelisensiering og rett til publisering.

Arbeidsgruppen legger til grunn at avklaring av rettigheter blir en viktig oppgave i forbindelse med innhenting av materiale til Språkbanken, og vi anbefaler at man allerede i etableringsfasen får utarbeidet standardavtaler for innhenting av språkressurser.

Arbeidsgruppen anbefaler for øvrig at staten ved framtidige bevilgninger til forskning på språkressurser stiller som vilkår at offentlig finansierte språkressurser skal stilles til rådighet for Språkbanken på samme vilkår som drøftet over.

Arbeidsgruppen har også merket seg at KKD skal påbegynne arbeidet med en overordnet språklov. I den forbindelse kan det være aktuelt å vurdere om språkloven skal inneholde regler av betydning for Språkbanken, så som for eksempel avleveringsplikt etter modell av lov om avleveringsplikt for allment tilgjengelige dokument. Framtidige, relevante ressurser som blir opparbeidet med offentlig støtte, f.eks. gjennom prosjekter finansiert av Forskningsrådet, bør ha en klausul om avleveringsplikt til Språkbanken.

6.2 Regler for statsstøtte

Språkbanken vil være en offentlig virksomhet som forventes å få en vesentlig del av sin finansiering fra staten.

Det offentliges økonomiske virksomhet er underlagt reglene om offentlig støtte på lik linje med private foretak. Det betyr at offentlig aktivitet i utgangspunktet ikke kan motta støtte i større utsterkning eller på andre vilkår enn privat virksomhet. Det offentliges tilskudd til Språkbanken må derfor vurderes i henhold til reglene om statsstøtte.

I henhold til EØS-avtalens art. 61 er offentlig støtte til næringslivet som hovedregel forbudt. Forbudet retter seg mot selektive støttetiltak til all form for økonomisk

virksomhet. Hva som er offentlig støtte i henhold til EØS-avtalen, tolkes vidt. Forbudet er imidlertid ikke absolutt, og det er vedtatt ulike regelverk med betingelser for når ulike støttetiltak kan være forenlige med EØS-avtalen.

Støtten kan falle utenfor regelverket fordi det dreier seg om ordinært tjenestekjøp hvor støttemottaker ikke mottar noen økonomisk fordel, eller med hjemmel i EØS-avtalens art. 59 (2) for tjenester av allmenn økonomisk betydning. Arbeidsgruppen antar at bevilgningene til Språkbanken kan karakteriseres som tilrettelegging av en tjeneste av allmenn økonomisk betydning, og at det offentliges finansiering av Språkbanken ikke omfattes av statsstøttereglene.

Dersom det skulle vise seg at Språkbankens virksomhet omfattes av statsstøttereglene, må det vurderes om det finnes lovlige unntak fra det generelle forbudet.

Generelle vilkår for lovlige unntak fra statsstøttereglene følger av EØS-avtalens art. 61 (3). Statsstøtte etter denne bestemmelsen kan tildeles til legitime formål, det vil si i tråd med EØS-avtalen. Det kan dreie seg om tiltak for å avhjelpe markedssvikt eller tiltak for å nå politiske og sosiale målsetninger som er forenlige med EØS-avtalen. Dernest må støtten være nødvendig for å oppnå de aktuelle målsetningene, og endelig må støtten veies opp mot de konkurranseskadelige virkningene, det såkalte proporsjonalitetsprinsippet.

Avgjørende for om statsstøttereglene kommer til anvendelse, er om støttemottaker er et foretak i EØS-avtalens forstand. Foretaksbegrepet omfatter som nevnt enhver juridisk enhet som driver økonomisk aktivitet, også offentlig virksomhet. Vurderingen av aktiviteten er uavhengig av enhetens juridiske status og måten virksomheten finansieres på.

Med økonomisk aktivitet menes en virksomhet som består i å tilby varer og/eller tjenester i et marked. Organisasjoner som alene ivaretar sosiale eller kulturelle funksjoner, vil som regel ikke anses for å drive økonomisk aktivitet. Grensedragningen mellom økonomisk og ikke-økonomisk aktivitet kan være vanskelig, og markedsforholdene i enkelte sektorer kan utvikle seg og endre seg over tid. Det bør vurderes nærmere om Språkbanken vil komme til å tilby tjenester i et marked. Til tross for at det er lagt til grunn i St.meld. nr. 35 at ingen andre enn Norge vil etablere en nasjonal språkbank, kan det ikke utelukkes at private aktører etablerer språkressurser som delvis kan være i konkurranse med Språkbankens ressurser.

For at offentlig støtte skal være forbudt, må støtten ha virkning på konkurransen og samhandelen innenfor EØS-området. I praksis vurderes ofte de to førstnevnte kriteriene samlet. Rettspraksis viser at det skal lite til før en anser et støttetiltak for å vri eller true med å vri komkurransen. For å kunne identifisere de konkurransemessige virkningene av offentlig støtte må man kunne identifisere det eller de relevante markedene der støttemottaker driver sin økonomiske aktivitet. Det anbefales som nevnt at det gjøres en enkel analyse av det relevante marked. Dette kan praktisk gjennomføres ved å ta utgangspunkt i hvilke produkter og tjenester Språkbanken skal tilby, og undersøke om det er alternative tilbydere av hele eller deler av dette produkt- og tjenestespekteret. En sammenligning med praksis i for eksempel ELDA kan være et utgangspunkt.

Reglene om stattsstøtte omfatter også forbud mot kryssubsidiering. Kryssubsidiering er særlig aktuelt for offentlig støtte til foretak som både utfører offentlige tjenester og konkurranseutsatt virksomhet, og det må vurderes om dette kan være aktuelt for Språkbanken. Kryssubsidiering mellom de forvaltningsmessige og kommersielle tjenestene kan medføre ulike konkurransevilkår for de kommersielle tjenestene for Språkbankens eventuelle konkurrenter.

Et virkemiddel mot kryssubsidiering kan være organisatoriske tiltak, det vil si at man skiller ut de ulike aktivitetene i separate rettssubjekter. Et juridisk skille i en så vidt begrenset virksomhet som Språkbanken er imidlertid ikke hensiktsmessig.

Et annet tiltak mot kryssubsidiering kan være at det opprettes egne regnskap for den kommersielle driften og for kjernevirksomheten. Den reelle kostnadsfordelingen kan da lettere sjekkes i etterkant. Arbeidsgruppen vil anbefale at en om nødvendig benytter dette virkemidlet.

6.3 Regler for offentlige anskaffelser

Uavhengig av organisasjonsform vil Språkbankens kjøp av varer og tjenester etter arbeidsgruppens oppfatning komme inn under reglene for offentlige anskaffelser. Språkbanken vil følgelig måtte konkurranseutsette både anskaffelse av språkressurser og kjøp av tjenester for eksempel av språkfaglig art i den utstrekning det er grunnlag for konkurranse.

Det bør vurderes nærmere i hvilken utstrekning de offentlige anskaffelsesreglene kommer til anvendelse på statlig finansierte språkressurser, og på identifiserte, unike språkressurser der man vet det kun finnes én utgave eller kun ett sted der den faglige kompetansen befinner seg.

Arbeidsgruppen har ikke tatt høyde for eventuelle merutgifter til gjennomføring av anskaffelsesprosesser i sine budsjettforslag.

6.4 Skatt og avgift

Språkbanken vil etter arbeidsgruppens anbefaling om organisasjonsform i utgangspunktet være et skattepliktig subjekt. Språkbanken vil imidlertid ikke ha erverv til formål og den vil dermed etter arbeidsgruppens oppfatning omfattes av unntaket fra skatteplikt i Skatteloven § 2-32. I den utstrekning Språkbanken driver økonomisk virksomhet, vil inntekten fra denne virksomheten være skattepliktig.

Det kan, i hvert fall de første årene, være grunnlag for momsfritak.

6.5 Personvern

Arbeidsgruppen legger til grunn den juridiske betenkningen fra 2002 hva gjelder personvernspørsmål. Gruppens konklusjon er at Språkbanken vil være behandlingsansvarlig, og at Språkbankens virksomhet vil være meldepliktig etter personopplysningsloven (pol). Konsesjonsplikt kan tenkes å være relevant for eventuelle opplysninger om etnisitet. Språkbanken vil ellers ikke behandle sensitive personopplysninger.

Arbeidsgruppen legger videre til grunn at innsamling av språkressurser må baseres på informert samtykke, og det bør utarbeides en samtykkeerklæring. Eventuelle tredjeparter som samler inn språkressurser på vegne av Språkbanken, må pålegges å benytte samme samtykkeerklæring. I de tilfeller Språkbanken innhenter kopi av eksisterende språkressurser, må det vurderes om det må innhentes særskilt samtykke for Språkbankens bruk.

Det har vært avholdt møte med Datatilsynet hvor arbeidsgruppens oppfatning ble foreløpig bekreftet. Datatilsynet anbefalte at det tidlig tas stilling til i hvilken utstrekning språkressursene kan anonymiseres. Datatilsynet vil være behjelpelig med gjennomgang av den samtykkeerklæringen som må utarbeides for innsamlingsformål, og eventuelle avtaler med tredjeparter om innsamling og bruk av språkressurser.

7 Økonomiske konsekvenser

Språkbankens administrasjon koordinerer arbeidet med innsamling, tilrettelegging og distribusjon av innhold, men mye av arbeidet utføres i fagmiljøene ved at man kjøper tjenester. Prisen på kjøp av tjenester fra fagmiljøene er beregnet i avsnittet om utvikling og drift av Språkbanken. Avklaring av rettigheter, også problemstillinger knyttet til personvern, vil være en del av administrasjonens løpende oppgaver.

7.1. Administrasjon

I innstillingen om norsk språkbank fra 2002 er det lagt til grunn at administrasjonen skal være liten. Arbeidsgruppen har gått gjennom rapporten fra 2002 og finner at behovet for personell ikke har endret seg. Det trengs en daglig leder med kunnskap om administrasjon og språkteknologi, en datalingvist og en ingeniør med høy it-kompetanse. Språkbanken vil trenge tjenester som juridisk bistand, regnskapsføring, revisjon osv., og disse må kjøpes etter behov. For eksempel vurderer arbeidsgruppen det som rimelig at det trengs et større volum på juridiske tjenester i starten enn når man har kommet i gjenge med arbeidet,

mens behovet for regnskapstjenester antas å være omtrent det samme over flere år. Utgiftene er beregnet i 2008-kroner og på tilsvarende grunnlag som i rapporten fra 2002, men justert til dagens lønns- og prisnivå. Tjenester som må kjøpes hos fagmiljøene når det gjelder innsamling, bearbeiding og tilrettelegging av språkressurser for Språkbanken, er lagt inn i budsjettet for drift og utvikling. Det er en forutsetning for en så liten administrasjon at personalet har med seg et nettverk og kan bygge videre på dette. Språkbanken vil være avhengig av tett og god kontakt med samarbeidsparter nasjonalt og må delta i nordiske og internasjonale nettverk (f.eks. CLARIN og FlaReNet) for å være tilstrekkelig oppdatert på området.

Følgende administrative oppgaver må ivaretas av Språkbanken:

- utarbeide generelle retningslinjer for ressurser til Språkbanken
- identifisere aktuelle språkressurser og leverandører
- avklare brukerrettigheter for nye ressurser
- gi informasjon om tilgjengelige ressurser (inkl. dokumentasjon)
- iverksette innsamlings- og kvalitetssikringsarbeid
- iverksette tilrettelegging som oppdatering av eksisterende ressurser, standardisering av materiale, oppmerking osv.
- behandle søknader om tilgang
- utarbeide og skrive kontrakter med brukere og leverandører
- være oppdatert på internasjonale standarder på området
- ivareta nasjonale og internasjonale samarbeidsrelasjoner
- kopiere og distribuere språkressurser
- utføre driftsoppgaver knyttet til distribusjon av språkressurser
- utføre generelle administrative oppgaver

Administrative kostnader for Språkbanken

	2009 a)	per år 2010–2014
Lønn, 3 årsv. inkl. sos.utg.	2 100 000	2 100 000
Husleie, kontorutg.	300 000	350 000
Møtegodtgjørelse	50 000	50 000
Konsulentbistand (jur o.a.)	200 000	150 000
Informasjon, nasjonale og internasjonale kontakter	150 000	100 000
Regnskap og revisjon	200 000	250 000
Samlet	3 000 000	3 000 000*

a) Engangskostnader til kjøp av utstyr etc. samt aksjekapital er ikke med i summene her.
*Det er beregnet en moderat prisøkning på husleie og regnskap/revisjon, men ikke noen lønnsøkning for årene 2010–2014.

7.2 Utvikling og drift

Mye av tilretteleggingen av ressurser som skal gjøres tilgjengelige gjennom Språkbanken, må gjøres av kvalifisert personell hos leverandørene av ressursene. De samme forutsetningene ble lagt til grunn for beregningene i 2002, og situasjonen er tilsvarende i dag. En del aktuelle språkressurser kan være godt merket og tilrettelagt for gjenbruk, andre har god basiskvalitet, men kan mangle oppmerking eller standardisering, og denne typen oppgaver må utføres ved at oppdrag gis til de aktuelle fagmiljøene. Oppgavene vil variere i omfang, og det vil også variere hvilken type fagkompetanse som er nødvendig. Disse tjenestene må kjøpes fra fagmiljøene. I tillegg kommer utgifter til validering og kvalitetskontroll som må gjøres av eksperter fra tilsvarende fagmiljø andre steder, i mange tilfeller betyr det internasjonalt.

I 2002-rapporten var gjennomsnittlige kostnader per årsverk anslått til kr 575 000. Vi har nå basert oss på kr 750 000 per årsverk. Dette svarer til en kostnadsøkning på om lag 30 %, på linje med den gjennomsnittlige lønnsutviklingen i staten i perioden. I summen har vi tatt høyde for at det trengs ulike typer fagekspertise til ulike oppgaver, og at vi kanskje må kjøpe tjenester også fra ikke-statlige organisasjoner der timeprisene kan være høyere enn i universitetsektoren.

Utgiftene til innsamlings- og utviklingsarbeid er beregnet ved at vi har vurdert antall årsverk og multiplisert med gjennomsnittlig kostnad per årsverk. Det er en forutsetning for Språkbankens organiseringsmodell med den lave bemanningen at denne type arbeid ikke utføres av Språkbankens eget personell, men settes bort til relevante fagmiljøer. Språkbanken skal opptre slik at den bidrar til å beholde og støtte de små fagmiljøene vi har på de ulike språkteknologiområdene i Norge, i stedet for å trekke fagekspertisen bort fra dem. Oppdrag fra Språkbanken vil i noen tilfeller også kunne bidra til å gi miljøene nye utfordringer og øke kompetansen der det finnes et godt grunnlag fra før.

Fra dag en vil Språkbanken kunne ha språkressursene etter NST som er tilrettelagt for distribusjon, nemlig taleressursene. Konsortiet som kjøpte boet etter NST, har en intensjonsavtale som tilsier at materialet skal stilles til rådighet for Språkbanken når den er etablert. Konsortiet har sørget for en gjennomgang av boet og har særlig konsentrert seg om de akustiske basene. Disse er tilrettelagt for gjenbruk, de er godt merket, og alle rettigheter er avklart. De kan trenge tillegg av nytt materiale når det gjelder representasjon av dialekter og spredning på alder blant informantene. Leksikonet som hører til, bør oppdateres med nytt materiale for å tilfredsstille kravet til "ferskvare". I tillegg er her rettigheter til deler av underlagsmaterialet som må avklares før det kan distribueres. Avklaring av rettighetene er en del av administrasjonens oppgaver i 2009. Med de forespørslene om tilgang til bruk av de akustiske basene og leksikonet som alt har kommet, vil det være rimelig å prioritere arbeidet med disse i 2009.

Språkbanken er primært opptatt av å kjøpe bruksrett til ressurser og verktøy som skal gjøres tilgjengelige gjennom Språkbanken. Det er i skrivende stund ikke avklart hvordan konsortiet som kjøpte boet etter NST, forholder seg til overføring av bare bruksretten til Språkbanken, eller om det ønsker at Språkbanken skal overta eiendomsretten. Konsortiets styre vil ta dette opp på første styremøtet høsten 2008.

Boet etter NST innholder talemateriale som er systematisk tatt opp og godt merket for språkteknologisk bruk. Dette betyr at Språkbanken får tilgang til et innhold som vi i 2002 så for oss måtte samles inn og bearbeides fra grunnen av. De akustiske basene utgjør ikke så stor prosentdel av en ressurssamling som Språkbanken skal være, men det er en type materiale som koster mye å bearbeide og tilrettelegge for gjenbruk. Vi vil anslå at vi sparer 10–15 årsverk i utviklingskostnader ved at Språkbanken får overta bruksretten til basene. I tillegg kommer besparelser pga. antatt høyere effektivitet i produksjonen som følge av bedre verktøy. Samlet har vi beregnet at vi nå trenger ca 40 årsverk mindre enn i 2002. Språkbanken må fortsatt belage seg på at bruksretten vil koste noe, at man trenger å legge til en del materiale for å oppdatere ressursene og at man må betale for kvalitetskontroll av materialet.

Tar vi utgangspunkt i dette anslaget for antall årsverk som trengs og korrigerer for lønnsutviklingen, får vi et beregnet behov på 10,5 mill. kr til kjøp av bruksretter og 74,5 mill. kr til nyutvikling. Arbeidsgruppen har kombinert denne grove beregningen med mer detaljerte vurderinger. De detaljerte tallene presenteres i tabellene 6,7 og 8.

			Minimum			Ønsket		
Type	Talestil	Formål	Kostnad, nyinnsamling	Kostnad, innkjøp	Totalkostnad, minimum	Kostnad, nyinnsamling	Kostnad, innkjøp	Totalkostnad, ønsket nivå
Romkvalitet	Spontan	Diktering, dialoger	12 863	1 838	14 701	33 075	1 838	34 913
Romkvalitet	Manuskript	Diktering, modeller	0	1 000	1 000	6 762	3 969	10 731
Telefon	Manuskript	Modeller	0	1 000	1 000	2 457	3 063	5 519
Mobiltlf	Manuskript	Modeller	0	1 470	1 470	588	1 470	2 058
Tlf i bil	Manuskript	Diverse	5 880	0	5 880	11 760	0	11 760
Telefon	Spontan	Dialog	3 675	0	3 675	7 350	0	7 350
Telefon	Spontan	Diktering	3 675	0	3 675	7 350	0	7 350
Romkvalitet	Manuskript	Difondatabase	990	0	990	1 980	0	1 980
Romkvalitet	Manuskript	Prosodi / Lydbibliotek	1 315	5	1 320	2 630	5	2 635
Kringkasting	Variert	Emnesøk	0	1 287	1 287	4 786	1 286	6 072
Telefon	Manuskript	Emnesøk i multimedia-arkiv	1 315	0	1 315	2 630	0	2 630
Audio	Spontan	Emnesøk	2 940	147	3 087	6 615	147	6 762
Romkvalitet	Spontan	Multimodale grensesnitt	0	551	551	2 581	551	3 132
Romkvalitet	Spontan	Møtetranskripsjon	0	0	0	3 675	0	3 675
Høy romkvalitet	Manuskript	Konkatenativ talesyntese	368	184	552	1 111	184	1 294
SUM			**33 021**	**7 481**	**40 502**	**95 349**	**12 512**	**107 861**

Tabell 6. Estimerte kostnader (i 1000 kr), taledata

		Minimum			Ønsket		
Teksttyper	Bearbeiding	Kostnad, nyinnsamling	Kostnad, innkjøp	Kostnad, minimum (bm & nn)	Kostnad, nyinnsamling	Kostnad, innkjøp	Kostnad, ønsket nivå (bm & nn)
Tospråklige tekster (norsk - engelsk)	Basal tilrettelegging	4 116	74	4 190	8 396	74	8 470
Tospråklige tekster (engelsk - norsk)	Basal tilrettelegging	4 116	74	4 190	8 396	74	8 470
Tospråklige tekster (engelsk - norsk og norsk - engelsk)	Grundig tilrettelegging	2 875	0	2 875	5 750	0	5 750
Sakprosa, småtrykk, upublisert materiale	Basal tilrettelegging	5 880	1 260	7 140	31 164	1 260	32 424
Aviser og media, skjønnlitteratur	Basal tilrettelegging	5 880	1 260	7 140	31 164	1 260	32 424
Sakprosa, småtrykk, upublisert materiale	Utvidet tekstkoding, manuelt kontrollert ordklassemerking	1 470	0	1 470	2 875	0	2 875
Aviser og media, skjønnlitteratur	Utvidet tekstkoding, manuelt kontrollert ordklassemerking	1 470	0	1 470	2 875	0	2 875
Aviser	Etablering av trebank	1 764	0	1 764	8 461	0	8 461
Anonymiserte journaler	Treningsdata for medisinsk diktering	0	0	0	458	0	458
SUM		**27 571**	**2 667**	**30 238**	**99 539**	**2 667**	**102 206**

Tabell 7. Estimerte kostnader (i 1000 kr), tekstdata

	Minimum			Ønsket nivå		
Aktivitetstype	Kostnad, nyinnsamling	Kostnad, innkjøp	Kostnad, minimum	Kostnad, nyinnsamling	Kostnad, innkjøp	Kostnad, ønsket nivå
Kjøp og kvalitetskontroll ordlistedata, bokmål	990	750	1 740	2 040	1 575	3 615
Kjøp og kvalitetskontroll ordlistedata, nynorsk	1 830	630	2 460	3 930	630	4 560
Innlemming av ordlister fra ulike kilder	2 630	0	2 630	2 630	0	2 630
Utvikling av stavevarianter/basis dialektvarianter	1 640	0	1 640	3 280	0	3 280
Utvikling av uttalebeskrivelse for navn, fremmedord og nyord	2 940	0	2 940	3 435	0	3 435
Uttalebeskrivelser for dialektregioner	2 940	0	2 940	3 590	0	3 590
Tospråklige parallellordlister	0	315	315	1 050	315	1 365
Ordnett	2 058	21	2 079	4 158	21	4 179
Begrepsontologi (SIMPLE)	1 680	210	1 890	3 780	210	3 990
SUM	**16 708**	**1 926**	**18 634**	**27 893**	**2 751**	**30 644**

Tabell 8. Estimerte kostnader (i 1000 kr), leksikalske ressurser

Summerer vi tallene i tabellene 6, 7 og 8, får vi et behov på 12,1 mill. kr til kjøp av bruksretter og 77,3 mill. kr til nyutvikling. Dette ligger svært nær den grovere beregning på hhv. 12 mill. kr og 78 mill. kr, og viser at det er god konsistens i tallgrunnlaget. Arbeidsgruppen holder fast ved anslaget om en ramme på på 90 mill. kr til investeringer. Oppstartsutgiftene i tillegg.

Vi må ta forbehold om at en grundigere gjennomgang (se forslaget til forberedende arbeid h-2008) kan bety justeringer på fordelingen på årene og fordelingen mellom kjøp av bruksretter og utvikling av nye ressurser.

Vurderingen tilsier at vi for de neste seks årene har behov for omlag 78 mill. kr til bearbeiding og tilrettelegging av ressurser og verktøy som Språkbanken må ha om vi følger anbefalingene i BLARK. BLARK-konseptet er nyttig når det gjelder å identifisere innhold i en språkbank, noe annet er en realistisk fordeling når det gjelder hvilke oppgaver vi kan regne med å få utført, og hvor de kan utføres. Her må man ta hensyn til ressurstilgang når det gjelder fagpersonalet i de ulike miljøene, og sørge for at belastningen blir fordelt slik at det vil tjene Språkbankens formål og fagmiljøenes. En satsing på språkteknologisk forskning av noe omfang fra Forskningsrådets side kan bety at innholdet i Språkbanken kan øke raskere enn vi ser for oss i dag, og det kan bety at vi kan få annet materiale. Stikkordet her er koordinering og samarbeid.

Setter vi dette inn i et samlet investeringsbudsjett (i mill. kr) for perioden, får vi:

	Adm.	Kjøp av bruksr.	Bearb./utv.	SAMLET
2009	3,0	2,5	1,5	(7,0+5) 12,0*
2010	3,0	3,5	14,5	21,0
2011	3,0	2,5	16,0	21,5
2012	3,0	1,5	17,0	21,5
2013	3,0	1,0	18,0	22,0
2014	3,0	1,0	11,0	15,0
SAMLET	18,0	12,0	78,0	113,0

* Et aksjeselskap må ha en aksjekapital. I tillegg må det avsettes midler til oppstartsutgifter (utrustning av kontor, anskaffelse av servere og annet datautstyr). Disse utgiftene, som vi har anslått til 5 mill. kr, foreslår vi bevilges i 2009, hvorav 3 mill. kr til aksjekapital. I tabellen framkommer oppstartutgiftene bare i kolonnen SAMLET.

8 Konklusjoner og anbefalinger

Norsk språkbank skal være et tiltak for å oppnå hovedmålet med stortingsmelding 35 (2007–08): "... språkpolitikk med det overordna målet å sikra det norske språkets status og bruk på alle samfunnsområde, slik at norsk kan bestå som eit fullverdig, samfunnsberande språk". Språkmeldingen setter også klare mål for en norsk språkbank: "Språkbanken skal vera ein stor, samla, nasjonal språkressurs som er kvalitetssikra og bygd opp etter internasjonale standardar."

Vår visjon for Språkbanken er at den skal være det naturlige samlingspunktet for lagring og distribusjon av offentlige og private digitale språkressurser. Ut fra dette foreslår arbeidsgruppen:

- Norsk språkbank blir etablert som aksjeselskap med eget styre fra 1.1.2009.

- Staten ved Kultur- og kirkedepartementet blir hovedeier i selskapet.

- En detaljert norsk BLARK-analyse legges til grunn for arbeidet med etableringen.

- Etableringsfasen for Norsk språkbank settes til 6 år, med en investeringsramme på 90 mill. Årlig administrativ ressurs er beregnet til 3 mill.

- Språkbankens formål er å lette tilgangen til eksisterende språkressurser og verktøy for språkteknologisk forskning og utvikling for både private og offentlige aktører. En forutsetning er også at de aktuelle miljøene bidrar til utvikling av Språkbankens ressurser.

- Språkbanken vil ha en liten administrasjon og bør lokaliseres i tilknytning til et relevant språkteknologisk eller språkfaglig miljø. Arbeidsgruppen mener at en samlokalisering med

Språkrådet kan være en god løsning. Det kan likevel tenkes andre løsninger for lokalisering, ut fra arbeidsgruppens prioriteringer.

- Språkbanksatsingen må følges opp og koordineres med en sterk satsing på språkteknologisk forskning.

- Arbeidsgruppen har diskutert sterke og svake sider ved den norske språkbanksatsingen i en egen analyse som er vedlagt rapporten (SVOT-analyse).

Referanser

Binnenpoorte, 2002: D. Binnenpoorte, F. de Vriend, J. Sturm, W. Daelemans, H. Strik, C. Cucchiarini: "A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch", Proc. LREC 2002, Las Palmas, 2002

Krauwer, 2006: S. Krauwer, B. Maegaard, K. Choukri, L. D. Jørgensen: "Report on BLARK for Arabic", http://www.nemlar.org/Publications/BLARK-final_190906.pdf, 2006

Hein, 2006: A. S. Hein, E. Forsbom: "A Swedish BLARK", GSLT Retreat Workshop, Gullmarstrand, http://stp.ling.uu.se/blark/sveblark060129.pdf, 2006

Samling og tilgjengeleggjering av norske språkteknologiressursar, Språkrådet 2002

Simonsen Føyen Advokatfirma DA, 2002: "Betenkning over juridiske problemstillinger knyttet til samling og tilgjengeliggjøring av norske språkteknologiressurser"

St.meld. nr. 48 (2002-2003) Kulturpolitikk fram mot 2014, Kulturdepartementet 2003

St.meld. nr. 35 (2007-2008) Mål og meining, Kulturdepartementet 2008

Vedlegg

Vedlegg A: Definisjonsdel av norsk BLARK-analyse

Vedlegg B: Oversikt over eksisterende språkressurser

Vedlegg C: SVOT-analyse for Språkbanken

Vedlegg D: Vedtekter for Norsk Språkbank

Vedlegg E: Arbeidsliste for oppstarten av Språkbanken

Vedlegg A: Definisjonsdel av norsk BLARK-analyse.

Definisjonsdelen av BLARK-analysen viser i hvilken grad de 14 identifiserte språkteknologisek anvendelsene er avhengige av spesifikke språkteknologiske moduler, og i hvilken grad de individuelle modulene er avhengige av ulike typer språkdata.

Analysen må betraktes som preliminær. Selv om et mindre antall sentrale fagpersoner har kommet med innspill og kommentarer, er denne versjonen basert på en stor grad av skjønnsmessig vurdering. En mer omfattende høringsrunde er nødvendig for å kunne etablere en kvalitetssikret, omforent BLARK-definisjon for norsk.

I presentasjonen har vi valgt å skille tekstbaserte moduler fra talebaserte moduler for å lette oversikten. Tabell A.1 viser hvordan de ulike anvendelser avhenger av taleteknologiske moduler. Tabell A.2 viser hvordan de ulike taleteknologiske modulene avhenger av dataressurser. Tabell A.3 og A.4 viser tilsvarende oversikter for tekstorienterte moduler.

Avhengigheten angis på en 4-punktsskala: Essensiell (+++), svært viktig (++), viktig (+), ikke viktig (' '). Dette samsvarer med ELRAs bruk av BLARK.

Modul/Anvendelse	Dataassistert språklæring	Adgangskontroll	Talestyring	Transkripsjon	Diktering	Tekst-til-tale	Dialogsystemer	Dokumentproduksjon	Automatisk sammendrag	Informasjonsgjenfinning	Informasjonsaksess	Oversetting (tekst-tekst)	Oversetting (tale-tale)
Komplett talegjenkjenner	++	++	+	+++	+++		+++	++	++	+++	++		+++
Akustiske modeller	+++	++	+++	+++	+++	++	++	++	++	+++	++		+++
Språkmodeller	+++	+	+	+++	+++	+	++	+++	+++	+++	++	+++	+++
Uttaleleksikon	+++	++	+	+++	+++	+++	++	+++	+++	+++	++		+++
Robust talegjenkjenning	++	+++	+++	++	+		+++	+	+	+	++		++
Talegjenkjenning for dialekter og innvandrernorsk		++	++	++	++	++	++	++	+	++	++		++
Taleradapsjon	+	+	++	++	+++	+	+	+	+	+	++		+++
Leksikalsk adapsjon	++	+	++	++	++	+	++	+	+	++	++		++
Prososdi-gjenkjenning	++	+	+	+	+	++	++	+	+	+	+		++
Komplett talesyntese	++	+				+++	++			+	+		+++
Difonsyntese	+	+				+	+			+	+		+
Skjøtesyntese	++	+				+++	++			+	+		+++
Prosodiprediksjon for TTS	++	+				++	++			+	+		++
Automatisk fonetisk transkripsjon	++	+			+	++	+			+	+		++
Automatisk fonetisk segmentering	++	+			+	++	+			+	+		++
Fonetiske likhetsmål	+					++	+						+
Talergjenkjenning		+++		++			+			++	+		+
Talersporing		++		++				+		++	+		+
Språkidentifkasjon	++	+		++	++		++			++	+		+++
Diälektidentifikasjon		+		++	+		++			+	+		++
Konfidensmål	+++	+++	+			+	+++	+		+	++		++
Ytringsverifikasjon	++	+	+			+	+++	+			+		
Emosjonsidentifikasjon			+	+			++						+
Taledeteksjon	+++	+	+	+	+++		+++			+	+		++

Tabell A.1: Anvendelsers avhengighet av data, tale

Modul/Data	Enspråklige leksika	Navneleksika	Flerspråklige leksika	Tesaruser	Ontologier, ordnett	Annotert tekst	Ikke-annotert tekst	Manuskriptlest tale	Spontan tale	Talt dialog++	Flerspråklig tekst	Multimedie/-modale korpora
Komplett talegjenkjenner	+++	+++	+			+	+++	+++	++	++	+	++
Akustiske modeller	+++	+++	++			+	+++	+++	+++	+++	+	+
Språkmodeller	++	++	+			++	+++	+	+	+	+	+
Uttaleleksikon	+++	+++	++			+	+	++	++	++	+	+
Robust talegjenkjenning	+	+				+	+	+++	+++	+++	+	+
Talegjenkjenning for dialekter og innvandrernorsk	+++	+++	+++			+	++	++	+++	+++	++	+
Taleradapsjon	+	+				+	+	+++	++	++	+	++
Leksikalsk adapsjon	+++	+++	+++			++		++	++	++	+	+
Prososdi-gjenkjenning	+	+	+			+++	+	++	++	++	+	+
Komplett talesyntese	+++	+++	+			++	+	+++				+
Difonsyntese	+++	+++	+			+		++				+
Skjøtesyntese	+++	+++	+			++	+	+++				+
Prosodiprediksjon for TTS	+++	+++	+			++		++	++	+		+
Automatisk fonetisk transkripsjon	+++	++	++			+		+++	++	++		+
Automatisk fonetisk segmentering	+++	++	++			+		+++	++	++		+
Fonetiske likhetsmål	+	+	+			+		+++	++	++	+	+
Talergjenkjenning	+	+	+			+		+++	+	++		+
Talersporing	+	+				+	+	++	++	+++		+
Språkidentifkasjon	+	+	+++			+	+	+++	++	++	+++	+
Dialektidentifikasjon	+	+	++			+	+	+++	++	++		+
Konfidensmål	+	+				+	+	++	+++	++		+
Ytringsverifikasjon	+	+				+	+	++	+++	++		+
Emosjonsidentifikasjon	+	+				+	+	++	++	++		+
Tāledeteksjon	+	+				+	+	++	+	++		++

Tabell A.2: Modulers avhengighet av data, tale

Modul/Anvendelse	Dataassistert språklæring	Adgangskontroll	Talestyring	Transkripsjon	Diktering	Tekst-til-tale	Dialogsystemer	Dokumentproduksjon	Automatisk sammendrag	Informasjonsgjenfinning	Informasjonsaksess	Oversetting (tekst-tekst)	Oversetting (tale-tale)
Grafem-til-fonemkonvertering	++					+++	+++	+			+		+++
Token-deteksjon	+		+	+	+	+++	+	+++	+++	+++	++	+++	+++
Deteksjon av setningsgrenser	+		+	++	++	+++	+	+++	+++	+++	++	+++	+++
Navnegjenkjenning	+		+	++	++	+++	+	+++	+++	++	++	+++	+++
Stavekorreksjon	+					++		+++	++	++		++	++
Lemmatisering	+			+	+	++	+	++	+++	++	+	+++	+++
Morfologisk analyse	+			+	+	++	+	++	+++	++	++	+++	+++
Morfologisk syntese	+					++	+	++	++			+++	+++
Ord-disambiguering	++			++	++	++	++	++	++	+	+	++	++
Parsere, grammatikker	+			++	++	++	++	++	++	+	++	++	++
Grunn parsing	+			+	+	++	++	++	++	+	++	++	++
Konstituent-gjenkjenning	+			++	++	++	++	++	++	+	++	++	++
Semantisk analyse	+		+	++	++	++	+++		+++	++	++	+++	+++
Referentanalyse	+			++	++		++	++	+	+	++	++	++
Pragmatisk analyse	+			++	++	+	++		++	+	+	++	++
Tekstgenerering	+					++	++	++	++			++	++
Språkgjenkjenning (tekst)						++	+	+	++	++		++	
Språkavhengig oversetting	++							+			+	+++	+++
Sidestilling av parallelltekst												++	++
POS-tagger	+		+	+	+	++	+	+	++	+	++	++	++
Term-ekstraksjon	+								+			++	++

Tabell A.3: Anvendelsers avhengighet av data, tekst

Modul/data	Enspråklige leksika	Navneleksika	Flerspråklige leksika	Tesaruser	Ontologier, ordnett	Annotert tekst	Ikke-annotert tekst	Manuskriptlest tale	Spontan tale	Talt dialog++	Flerspråklig tekst	Multimedie/-modale korpora
Grafem-til-fonemkonvertering	++	++				++						
Token-deteksjon	+++	++				+	++					
Deteksjon av setningsgrenser	+	+				++	++					
Navnegjenkjenning	+++	+++	+	+	+++	++	++	++				
Stavekorreksjon	+++	+										
Lemmatisering	+++				+++	++	++					
Morfologisk analyse	+++				+++	++	+					
Morfologisk syntese	+++				+++	++	+					
Ord-disambiguering	+++	+			+++	++	+					
Parsere, grammatikker	+++	++			+++	++	+					
Grunn parsing	+++	++			+++	++						
Konstituent-gjenkjenning	+++				+++	++						
Semantisk analyse	+++	++		++	+++	++						++
Referentanalyse	+++	++		++	+++	++	+					++
Pragmatisk analyse	+++	+		+	+++	++						++
Tekstgenerering	+++	++		++	+++	++						++
Språkgjenkjenning (tekst)	++	+	++			+	++				++	
Språkavhengig oversetting	+++	+++	+++	++	+++	++		+			+++	
Sidestilling av parallelltekst	++	++	++			++	++				+++	
POS-tagger	+++	++			+++	+++	+					
Term-ekstraksjon	++	++				++	+					

Tabell A.4: Modulers avhengighet av data, tekst

Vedlegg B. Oversikt over eksisterende språkressurser.

I rapporten "Samling og tilgjengeleggjering av norske språkteknologiressursar" fra 2002 ble det presentert en oversikt over eksisterende materiale som burde vurderes i forbindelse med språkbanken (vedlegg 1 i rapporten fra 2002). I kartleggingen av eksisterende språkressurser har vi tatt utgangspunkt i i denne oversikten. Sommeren 2008 ble det i tillegg gjennomført en undersøkelse der et større antall relevante institusjoner ble bedt om å rapportere inn nytt tilfang av språkressurser siden 2002. De fleste av de forespurte institusjonene har gitt respons, selv om det på grunn av sommer og ferietid må påregnes at det kan være ressurser som ikke er fanget opp.

I denne informasjonsinnsamlingen har vi i tillegg til rene dataressurser også forsøkt å innhente informasjon om relevante språkteknologiske verktøy (jfr. BLARK-beskrivelsen i kapittel 3).

Det meste av det språkteknologisk relevante datamaterialet vil være materiale som avspeiler moderne norsk språkbruk. Vi har i tillegg fått inn informasjon om en del eldre språkdata. Dette er presentert i en egen tabell.

På de etterfølgende sidene er det gitt en tabellarisk oversikt over eksisterende materiale som kan tenkes å inngå i Språkbanken.

Tabell B.1: Taledata – oversikt

Type	Leverandør	Verdi	Tilgjengelig?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
Manuskript, 16 kHz	Språkbankkonsortiet (NST-data)	1,0 mill.	Alle rettar		Bokmål	540 t	Kan inngå
Manuskript, 8 kHz, telefon	Språkbankkonsortiet (NST-data)	1,0 mill.	Alle rettar		Bokmål	500 t	Kan inngå
Manuskript, 8kHz, mobiltelefon	Språkbankkonsortiet (NST-data)	1,0 mill	Alle rettar		Bokmål	200t	Kan inngå
TABU.0, Manuskript, 8 kHz telefon fastnett	Telenor FOU	1 mill.	Alle rettar		Bokmål og nynorsk	70t	Kan inngå
SpeechDat, manuskript	Telenor FOU	0,8 mill.	Alle rettar		Bokmål og nynorsk	60t	Kan inngå
Prosdata	Telenor FOU		Alle rettar		Bokmål	30 min	Prosodidatabase, kan inngå
SpeechDat Mobil	Telenor Mobil		Alle rettar			?	Kan inngå
Spontan tale, telefon	Telenor Mobil		?		Dialekt	16t?	2 CD-ar med opptak frå tenesta "Bare spør 1999". Må spesifiserast nærare.
Spontan tale, telefon	Telenor Business Solutions		?		Dialekt	?	Opptak frå automatisk sentralbordteneste: "Talk2Call". Må spesifiserast nærare.
Spontan tale, telefon	Telenor Kundeservice		?		Dialekt	~100.000 transkriberte samtaler	Opptak frå 145 og 05000 (feilmottak i Telenor der talegjenkjenning er innført). Må spesifiserast nærare.
Opplesen TV-teksting	SINTEF / NRK		Må avklarast			17 timar	Laga med tanke på å generere og teste akustiske modellar for automatisk teksting (gjenkjenning) av nyheitsprogram. Noko spontan tale. Må spesifiserast nærare.
Rundkast, nyhetssendinger, radio	NTNU/NRK		Forskning, annen bruk må avklares		Bokmål / nynorsk/ dialekt	75t	Transkriberte nyhetssendinger. - 30 min er fonetisk oppmerket. Rettighetsspørsmål må avklares pga musikkinnslag.
Radiosendingar	Nasjonalbiblioteket		Forsking			Meir enn 1500 timar	Lineær (48KHz,16bit), 384 kbps mpeg layer 2, 40 kbps mp3. Bør vurderast nærare.
FonDat, talesyntese	NTNU		FoU, annen bruk må avklares		Bokmål	2 talere á 8 t, 2 talere á 2t, 16 talere á 30min	Manuskriptlest tale for bruk i datadrevet skjøtesyntese utviklet av FONEMA-prosjektet.
Big Brother-korpuset - spontan tale	Tekstlab, UiO	Må avklares	Berre FoU	Må avklares	?	550.000 ord	Transkripsjon frå "reality"-serien. Spontantale. Bør kunne inngå
Manuskript, medisinske journaler	Max Manus	Må avklares	Alle rettar	Må avklares	Bokmål	200t	200 talere. Innsamlet for medisinsk diktering.

Type	Leverandør	Verdi	Tilgjengeleg?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
Digitaliserte lydbandopptak frå arkivet til INL.	Institutt for nordistikk og litteraturvitenskap, NTNU	0,8 mill. (OK kval?)	Må avklarast			15 timar, 20 min. transkr.	Dialektprøver. Bør vurderast for full transkripsjon og innleming for å oppnå bedre dialektdekning av taledataene
Stortingstaler / innlegg	Stortingets administrasjon		Alle rettar		Bokmål/nyn.	Ukjent	Alle stortingsrepresentantar sidan 1989. Må spesifiserast nærare.
LICHEN, kvensk	Tekstlab, UiO.	?	LICHEN-prosjektet	?	Kvensk	~70 intervjuer	
NoTa-Oslo, spontantale	Tekstlab., UiO	?	FoU og undervisning, annen bruk må avklares	Må avklares	Oslo-dialekt	~900.000 ord, 166 informanter	XML-kodet og grammatisk tagget med NoTa-taggeren. Ortografisk transkribert tale med lyd- og videofiler. Bør inngå.
Nordisk dialektkorpus	Tekstlab, UiO	?	FoU	Må avklares	Norske dialekter	400 informanter, 20 min samtale og 10 min intervju pr informant	Norsk del av nordisk prosjekt. XML-kodet, fonetisk transkripsjon og grammatisk tagging.
TAUS - intervjuer	Tekstlab., UiO	?	FoU	Må avklares	Oslo-dialekt	212.000 ord, 59 informanter	Ortografisk transkripsjon, XML-kodet og grammatisk tagget.
UPUS, dialoger	Tekstlab, UiO	?	UPUS-prosjektet, prosjektintern bruk	Må avklares	Multi-etnisk ungdomsspråk, Oslo	Intervju og samtaler med 55 ungdommer	Ortografisk transkripsjon koplet til lyd og videofiler. XML-kodet og grammatisk tagget.
Målføresamling	LLE, UIB	?	Forskning	Må avklares	Vestnorske dialekter	71t	13t er transkribert, ytterligere 35t under arbeid.
Målføresamling	Nordisk institutt, UiB		Må avklares		Dialekter, hovedsakelig vestnorske	1500t	Lydbånd.
Norsk tonelagstypologi	UiB		Kan inngå			15644 lydfiler	Testord i 138 konstruerte setninger
Digitale videoopptak, autentiske samtalar frå	Høgskolen i Agder		Må avklarast		?	8 timar	Deler av materialet er grovtranskribert. For det meste ortografisk transkripsjon

Tabell B.2: Tekstdata - oversikt

Type	Leverandør	Verdi	Tilgjengeleg?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
Tekst, ymse	Språkbankkonsortiet (NST-data)	0,5 mill	Avgrensa		Bokmål	~750 mill ord	Må undersøkes nærmere. Kan trolig benyttes til generering av statistiske språkmodeller. Annen bruk hemmet av opphavsrettslige årsaker.
Tekst, medisin	Språkbankkonsortiet (NST-data)	0,5 mill	Uklart		?	?	Må undersøkes nærmere
Tekst, medisin	Max Manus/Philips	Må avklares	Må avklares		Bokmål	414 mill ord	Journaltekst fra somatisk medisin, ca 50% radiologi
Skjønnlitteratur og faglitteratur	Samlaget		Uklart		Nynorsk	?	Svært aktuelt pga. behovet for mer nynorsk tekst i språksamlinga
Tekst for opplesing	NRK		Uklart		Nynorsk og bokmål	Uklart, men stort	Nynorskmaterialet er svært aktuelt
Avistekster	Aftenposten		Alle retter		Moderat bokmål	?	Har levert tekster til Tekstlab ved UiO - kvalitet kan vurderes derfra - bør kunne inngå
Diverse tekster	Leksikografisk avdeling ved UiO		Bare forskning		Bokmål	40 mill.	Skjønnlitteratur, sakprosa (rapportar/juridiske dokument, fagbøker, artikkelsamlingar, biografi), aviser, tidsskrift, ukeblad, noko TV-teksting, noko upublisert materiale (minneoppgåver, etnologi; ungdomsspråk fra prategrupper på nettet). Typen materiale .
Avistekster	HIT - UiB	0,5 mill.	Må avklares med avisene		Begge?	4,5 mill.	Eldre avistekster fra Aftenposten og Bergens Tidende.
Norsk aviskorpus	Aksis		Må avklares med avisene		Begge	670 mill ord	Nettbasert aviskorpus. Mest bokmål men noe nynorsk. Tagget med Oslo-Bergen-taggeren.
Oslokorpus, tagget bokmålstekst	UiO, Tekstlab	1,9 mill.	Bare forskning		Bokmål	18,3 mill.	Aviser (inkl Aftenposten nevnt annet sted), ukeblad, skjønnlitteratur, off dokument. Maskinelt tagga materiale. Klar kandidat dersom materialet også kan benyttes kommersielt.
Oslokorpus, tagget nynorsktekst	UiO, Tekstlab		FoU		Nynorsk	3,8 mill ord	Grammatisk tagget
Usenet	UiO, Tekstlab		Fritt tilgjengelig			140 mill ord	Tekster fra no-domenet av usenet i perioden 1998-2002. Kan trolig distribueres via Språkbanken.
KAL - elevtekster	UiO, Tekstlab, KAL-prosjektet		Forsning, utredning og opplæring			3300 elevtekster	
KB-N tekstbase	NHH / UiB		Forskning, annen bruk må avklares		Bokmål / engelsk	~450 000 ord parallelltekst, ~400.000 ord sammenlignbar tekst	Parallelle og sammenlignbare tekster på norsk og engelsk innen økonomisk-administrativt domene. XML/TEI-merket
NP-annotert norsk korpus	NTNU		Må avklares			Under utarbeidelse	
Div oversatte tekster	Lionbridge Norway	2,4 mill., dersom OK kval.	Må avklares med kunder	Data-kompensasjon	Bokmål	3,7 mill.	Ulike typer tekster som er oversatt til norsk. Klart relevant maskinoversettelse.
Parallelle fagboktekster	Fagbokforlaget		Må avklares	Datakompensasjon	Bokmål og nynorsk	870.000 ord	Bør kunne inngå.
Offisielle tekster, klassifisert når det gjeld målform v.h.a. spesialprogram	Fagbokforlaget		Alle retter	Datakompensasjon		250 MB, tilsvarar ?? ord	Må spesifiseres nærmere.
Div tekster	Aschehoug Videregående		Avgrensa retter	Datakompensasjon	Bokmål og nynorsk	"Svært stort"	Vanskeleg å vurdere.

Type	Leverandør	Verdi	Tilgjengeleg?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
Tekst og lydoptak	Nasjonalbiblioteket		Forskingsformål, noko må avklarast	Datakompensasjon	Bokmål, riksmål	Ca. 1 mill.	Stortingsreferat og memoarlitteratur. Kan vurderes.
Rapporter, skrift og temahefte	Norsk institutt for forsking om oppvekst, velferd og aldring		Forskingsformål	Datakompensasjon	Bokmål	?	Kan følges opp.
Manuelt tagga korpus bygget på avistekster	Lingvistisk institutt, NTNU	0,06 mill.	Alle retter	Datakompensasjon	Bokmål	60.000 ord	Stilles fritt til rådighet og bør inngå.
Tospråklig liste bokmål -> nynorsk.	Nynodata		Avgrensa rettar	Datakompensasjon	Bokmål / nynorsk	?	Bør kunne inngå dersom rettane blir avklart.
Div. tekst	Norge.no		Alle retter	Datakompensasjon	Bokmål / nynorsk	Ca. 3 mill.	Bør kunne inngå.
Innlegg på Stortinget	Stortingets administrasjon		Alle retter		Bokmål/nynorsk	Ukjent	Alle stortingsrepresentanter siden 1989. Må spesifiseres nærmere.
CELEX-korpuset	Aksis		Kan inngå		Bokmål / nynorsk	5000 tekster; 15 mill. ord	Norsk versjon av Eus dokumentdatabase.
ASK - Norsk andrespråkskorpus	UiB		Må avklares			100 tekster per morsmål, 10 ulike morsmål, også noe tilleggsmateriale, totalt ca. 750 000 ord	Tekstkorpus for norsk som andrespråk, typologisk ulike morsmål, samt norsk referansekorpus; grunnlag for kvantitative metoder i andrespråksforskning og pedagogisk utviklingsarbeid
Importord i Norden - korpus	UiB		Må avklares				Korpus med avistekster brukt for manuell ekserpering av nyord
Samiske tekster	UiT / Sametinget		Alle retter		Nordl-, sør- og lulesamisk	~8 mill ord	Morfologisk og syntaktisk annotert
Parallelltekster for inntil 30 språk	Oracle	Dersom 2 mill. ord, 1,2 mill. kr	Alle retter		?	20 mill. ord - uklart hvor mye som er relevant for norsk	Uklar fordeling norsk / engelsk. Verdi avhenger av hvordan materialet er kopla sammen
Parallelltekster norsk - engelsk	Universitetet i Oslo, Institutt for britiske og amerikanske studar (IBA)		Bare forskning ved IBA			2 mill.	Tekstene er koda i samsvar med TEI og tagga med ordklassemformasjon. Lemmatisert. Både original og oversatt tekst. Uegna utan endra bruksrett.
LOGON Tourist Corpus, norsk-engelske parallelltekster	Tekstlaboratoriet, UIO		FoU		Bokmål engelsk	175.000 ord	om
Oslo Multilingual corpus, flerspråklige parallelltekster	Tekstlaboratoriet, UIO		FoU ved UiO og UiB			15,5 mill ord	SGML-kodet og grammatisk tagget. Opphavsrettslige begrensninger som kan utelukke samlingen fra Språkbanken.
KIAP, parallellkorpus	UiB		Må avklares			450 artikler	Parallellkorpus med fagtekster, norsk, fransk, engelsk; økonomi/administrasjon, lingvistikk, medisin
Samisk-norske parallelltekster	Senter for samisk språkteknologi, UiT		Alle retter		Nordsamisk / bokmål	1,6 mill ord	Grammatisk tagget
Opus - flerspråklig	Tekstlaboratoriet, UIO		Fritt tilgjengelig		60 språk	30 mill ord	Kan trolig distribueres av Språkbanken
Engelsk-spansk parallellkorpus - CRISTINA	Aksis						
Engelsk-spansk parallellkorpus - Marlen	Aksis						
Engelsk-tysk parallellkorpus	Aksis						
Bosnisk tekst	Tekstlaboratoriet, UIO		FoU		Bosnisk	1,5 mill ord	

Tabell B.3: Leksikalske ressurser – oversikt

Type	Leverandør	Verdi	Tilgjengelig?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
Korpusgenererte leksikalske data	Språkbankkonsortiet (NST-data)	1 mill.	Alle retter	?	Bokmål	785.000. fullformer	Kan inngå. Transkribert. Overlapping med annet materiale fra universitetene
Kjerneleksikon	Språkbankkonsortiet (NST-data)	0,5 mill	Alle retter	?	?	230 000 ord, fonetisk transkribert	Kan inngå - overlapping med annet materiale fra universitetene
Leksika for flere språk	Oracle		Alle retter			14 mill. ord og fraser for flere språk	Kan vurderes, men må sjekkes for grammatisk oppmerking
Monolinguale leksika for ulike domener (og språk)	Kunnskapsforlaget		Alle retter			?	Bør inngå i den grad de ikke overlapper med annet materiale (f.eks. Fremmedord-boka, Medisinsk leksikon....). Må tilarbeides for infleksjon
Ordlister	Samlaget		Alle retter	?	Nynorsk	60-70 000	Kan komplettere NorKompLeks - bør inngå
Bokmål/nynorsk-ordlister	Samlaget		Alle retter	?		?	Høyaktuelt for bl.a. dokumentproduksjon på begge målformer
ONOMASTICA	Telenor		Alle retter		?	562 000 personnavn, bedriftsnavn osv. på 11 språk	Utdrag av relevante norske data er av stor interesse. Eierdel for Telenor? Også tilgjengelig via andre kanaler (ELRA)?? 50.000 for norsk?
Ordlister	NTNU - Telenor		Alle retter		Bokmål	Ca. 280 000 fullformer, grammatisk merkte og transkriberte	NorKompLeks. Kan inngå i samlinga om lag vederlagsfritt (avhenger av Telenor)
Ordlister	NTNU - Telenor		Alle retter		Nynorsk	Ca. 350 000 fullformer, grammatisk merkte og transkriberte	NorKompLeks. Kan inngå i samlinga om lag vederlagsfritt (avhenger av Telenor)
Ordlister generert fra småtrykk	Lionbridge Norway		Må avklares	Datakompen-sasjon	Bokmål og nynorsk	Ca 1 mill ord	Kan vurderes
Terminologi-ordlister for norsk/svensk/russisk/ tysk/engelsk/fransk	Fagbokforlaget		Alle retter	Datakompensasjon	Bokmål	Vel 900.000 ord	Overtatt fra Rådet for teknisk terminologi. Bør inngå.
Ordliste	Aksis		Må avklares	Datakompensasjon	Bokmål	74 000 lemmaer, 361 000 ordformer	Basert på prosjektet SCARRIE, som bygger på NorKompLeks. Inneholder svært nyttig stilinformasjon. Bør inngå
NOT: Ordliste over oljeterminologi - norsk/engelsk	Aksis		Alle rettar	Datakompensasjon	Bokmål, nynorsk	30 000 definisjoner	Vekt på oljeterminologi. Termer, synonym, kildehenvisinger, dels definisjoner. Kan inngå.
Ordlister, ordbøker og leksika	Kunnskapsforlaget		Avgrensa retter	Datakompensasjon	Bokmål, nynorsk	Uklart, men mye	Kunnskapsforlagets ordbøker. Av spesiell interesse for språkteknologi er synonymordbøker og tospråklige ordbøker
Ordlister	Universitetet i Oslo	4 mill. (behov 3,3)	Alle retter		Bokmål / nynorsk	110.000 1,2 mill. bøyde former	Ordsamlingene ved Dokumentasjonsprosjektet. Bør inngå
Nynorskordboka	UiO / Samlaget				Nynorsk	90.000 oppslagsord	
Bokmålsordboka	UiO / Kunnskapsforlaget				Bokmål	65.000 oppslagsord	
Norsk ordbok	UiO / Kunnskapsforlaget				Bokmål/riksmål	81.000 oppslagsord	

Type	Leverandør	Verdi	Tilgjengelig?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
Norsk Ordbok	UiO / Samlaget/Norsk Ordbok				Nynorsk	~300.000 artikler, 175.000 pr 2008	
Norsk ordbank	EDD, UiO		GPL-lisens/salg		150.000 oppslagsord for bokmål, 124.000 for nynorsk		Grunnformer fra Bokmålsordboka, Nynorskordboka, IBM-ordlista med mer. Grunnform pluss alle bøyningsformer.
Norsk Simple-database	UiO		Forskning		Bokmål	10.00 ordbetydninger	Norsk semantisk leksikon.
Nyordsboka 1975-2005	UiO		Nettordbok		Bokmål / nynorsk	3.000 artikler	
Norsk landbruksordbok	Landbruksmållaget				Nynorsk	~35.000 oppslagsord	
KB-N termbase	NHH / UiB		Forskning, annen bruk må avklares		Bokmål / engelsk	~8.500 poster, ~20.000 oppslagsord	Termbase på norsk og engelsk innen økonomisk-administrativt domene. RTT-standard.
Samisk termbase, flerspråklig	Sametinget		Alle rettigheter		Nord-, sør-, lule-samisk, norsk, svensk, finsk		
ScanlLex-leksikon, parallellordliste	Tekstlaboratoriet, UiO		Fritt tilgjengelig		Bokmål nynorsk, dansk, svensk, islandsk, engelsk	76.000 ordpar	Kan trolig distribueres via Språkbanken
Ordliste med animate substantiver	Tekstlaboratoriet, UiO		Fritt tilgjengelig				Kan trolig distribueres via Språkbanken
Importord i Norden	UiB		Kan inngå				Ordlister over nyere importord; også noe tekst
LEXIN	Utdanningsdirektoratet/Aksis		Må avklares			25.600 begrepsposter	Flerspråklige ordbøker for innvandrere. Oppmerking: Definisjoner, morfosyntaks, synonymi etc.
Norsk aviskorpus - nyorddatabase	Aksis		Må avklares		Bokmål / engelsk + 6 andre språk		Under oppbygging; manuelt kontrollerte poster vil inngå i Norsk ordbank. Planlagt annotering av morfosyntaks og orddannings mønstre (2009)
EU/E0S-basen	Aksis		Kan inngå		Bokmål /nynorsk / engelsk /fransk	40.000 begrepsposter, 130.00 enkeltemner	Termbase med søkbare termer på fransk, engelsk, bokmål og nynorsk fra over 70 fagområder brukt i oversettelse av EU-rettsakter
SCARRIE - leksikon	UiB		Kan inngå				Basert på NorKompLeks/Bokmålsordboka
Stadnamnsamlinga	Nordisk institutt, UiB		Kan inngå			200 000 navn	Navnedatabase, Vestlandet. Navn med delvis kobling til lydfiler.
TREPIL trebank	UiB		Kan inngå		Bokmål		Pilotversjon av en LFG-trebank for norsk bokmål

Tabell B.4: Verktøy – oversikt

Type	Leverandør	Årsverk	Tilgjengeleg?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
Anaforløsningssystem	Tekstlaboratoriet, UiO		Fritt tilgjengelig				Verktøy som finner anteedentene til pronominale anaforer i norske tekster. Kan trolig distribueres via Språkbanken
Norsk navnegjenkjenner	Tekstlaboratoriet, UiO		Fritt tilgjengelig				Del av Oslo-Bergen-taggeren. Lingvistiske regler kan trolig distribueres via Språkbanken
Norsk navnegjenkjenner 2	Tekstlaboratoriet, UiO		Ikke-kommersiell bruk med GPL-liseas				Statistisk navnegjenkjennenr som benytter Oslo-Bergen-taggeren som preprosessor. Kan trolig distribueres via Språkbanken
NoTa-taggeren	Tekstlaboratoriet, UiO		Fritt tilgjengelig				Statistisk talemålstagger. Kan trolig distribueres via Språkbanken
Osol-Bergen-taggeren	Tekstlaboratoriet, UiO		Ikke-kommersiell bruk med GPL-liseas		Bokmål / nynorsk		CG-basert morfologisk/syntaktisk tagger, flere underteknologier, inkl egennavngjenkjenning, kompositumanalyse; videreutvikles i aviskorpusprosjekte. Kan trolig distribueres via Språkbanken
ScanDiaSyns dialekttransliterator	Tekstlaboratoriet, UiO		Fritt tilgjengelig				Halvautomatisk dialektoversetter (dialekt til bokmål). Kan trolig distribueres via Språkbanken
Glossa	Tekstlaboratoriet, UiO		Fritt tilgjengelig				Webbasert verktøy for korpussøk. Kan trolig distribueres via Språkbanken
SIMPLE-redigering	Tekstlaboratorie og bokmåls-leksikografi, UiO		Fritt tilgjengelig				Redigeringssystem for norsk SIMPLE-ordbok (semantisk leksikon). Kan trolig distribueres via Språkbanken
Disambiguator for PP-tilordningen på norsk	Tekstlaboratoriet, UiO		Fritt tilgjengelig				Kan trolig distribueres via Språkbanken
TypeCraft	NTNU		Forskning		Flerspråklig		Verktøy og database for dyp annotasjon av tekster av liten og medium størrelse.
KB-N termekstraksjon	NHH / UiB		Forskning, annen bruk må avklares		Bokmål		Dataverktøy for automatisk termekstraksjon fra norsk tekst.
ASK - verktøy	UiB		Kan inngå				Annoteringsverktøy og søkesystem; mulig å søke etter feilkategorier, ord, lemma, strenger av ord, ordklasser og ulike kombinasjoner av disse
BRE:DT-ressursene	UiB		Kan inngå				Behandling av refererende enheter i diskursteori - System for automatisk merking av referansekjeder, anaforrelasjoner osv.

Type	Leverandør	Årsverk	Tilgjengeleg?	Kompensasjon	Målform	Omfang	Vurdering / kommentar
LFG ParserBanker	UiB		Kan inngå				Dataverktøy for oppbygging og av LFG-trebanker for norsk og andre språk, utviklet av TREPIL-prosjektet.
XLE-Web	UiB		Kan inngå				Web-grensesnitt til XLE for parsing med LFG-grammatikker, utviklet i LOGON-prosjektet
NorGram	UiB		Kan inngå		Bokmål		LFG-grammatikk for norsk
Norsk aviskorpus - verktøy	Aksis		Kan inngå				Verktøy for oppbygging og oppmerking av dynamisk korpus og leksikalsk database, bokmål/nynorsk-klassifikator, anglisismeidentifikasjon med mer
Ordnett-ressursene	UiB		Kan inngå				Verktøy for parallellkorpusbasert oppbygging av ordnett vha semantiske speil
SCARRIE - verktøy	UiB		Kan inngå				Sammensetningsgrammatikk, en setningsgrammatikk, grafem-til-fonemoversettelse
TCA2 - Text Corpus Aligner	Aksis		Kan inngå				Verktøy for setningsbasert parallellstilling (alignment) av parallelle tekster (automatisk/manuell kontroll)
GREI - grammatikkspill/trebank	VISL-prosjektet		Analyseresultater fritt tilgjengelig.		Bokmål / nynorsk		VISL-prosjektet ved Syddansk universitet har rettigheter til spill og andre verktøy. Kan benyttes til bygging av trebanker.
FONEMA-verktøy	NTNU		Alle rettigheter		Bokmål		Verktøy og prosedyrer for opptak og oppmerking av taledatabaser for skjotesyntese.
SVoG-verktøy	NTNU		Alle rettigheter		Bokmål		Verktøy og prosedyrer for opptrening av storvokabular talegjenkjenner for norsk. Akustiske modeller og statistiske språkmodeller kan inngå.

Tabell B.5: Samlinger med eldre språkdata

Navn	Ressurstype	Leverandør	Verdi	Tilgang	Målform	Omfang	Kommentar
Setelarkivet	Papirparkiv og database	informantar	uvurderleg originaltilfang	på nettet	målføre og nynorsk (1938)	3.03 mill setlar	Samling frå 1930 – 2001 digitalisert med faksimile , frå 2002 berre elektroniske setlar
Ordbokshotellet	database	forfattarar	Normering og indeksering ved ILN utgjer verdiauke	for NO-redaksjonen	målføre og nynorsk (1938)	28 ordsamlingar, 79094 postar	Starta 2006, digitaliserte ordbøker over norske målføre, under stadig vekst
Trønderarkivet	database	informantar	uvurderleg originaltilfang	på nettet	målføre og nynorsk	192203 postar	Digitalisert med faksimile
Det nynorske korpuset	tekst	norske forfattarar og forlag	digitalisering og tekstprosessering ?	på nettet	Nynorsk (Nynorsk rettskriving frå 1866 til i dag som i teksten, med hovudtyngd etter 1975.)	35 mill. ord	Monitorkorpus under stadig utviding. Inneheld fleire sentrale mest fullstendige forfattarskap og mykje vanskeleg tilgjengeleg tekst (t.d. fleire årgangar av Fedraheimen).
NO 2014 ordboksdatabasen	databasesystem	Norsk Ordbok	ca 280 årsverk	for NO-redaksjonen. NO 2014 er i produksjon, men nettutgåve er planlagd	målføre og nynorsk. 1938-rettskriving med somme modifikasjonar.	ca 25000 artiklar pr band, 175 000 ferdige artiklar per dato, tilsv 11200 boksider (1 spalte = 1 vanleg bokside)	NO 2014 skal presentere "portrett" i fullskala av ca 300 000 norske lemma. Om lag to tredelar har ikkje fått slik presentasjon før i noko norsk ordboksverk. Dette gjeld særleg for ordtilfanget frå norske målføre.
Grunnmanuskriptet	database	Norsk Ordbok	(prøvt påreikna) 30 årsverk	på nettet	Nynorsk (1917)	ca 13500 maskinskrivne sider, 105 000 artiklar	1917-rettskrivinga. Papirutgåve i eitt eksemplar i oppvaring hos NO 2014
Grunnmanuskriptet med blyantrettingar	database	Norsk Ordbok	2 årsverk (innføring av tillegg, digitalisering)	på nettet	Nynorsk (1917)	ca 110 000 artiklar	1917-rettskrivinga. Papirutgåve i eitt eksemplar i oppvaring hos NO 2014
Metaordboka	databasesystem	Norsk Ordbok	15 årsverk 2000 - 2002 + laupande vedlikehald	på nettet	Nynorsk (1938)	om lag 565000 artiklar	finst berre digitalt. 1938-rettskrivinga md somme modifikasjonar
Målføresynopsen	Digitaliserte bilete	Målførearkivet, UiO	papirutgåve - uvurderleg originaltilfang + digitalisering 1 årsverk	på nettet	Målføre innskrivi, og nynorsk i indeks		Indeks etter Storm, dvs rettskrivinga I Aasens ordbok 1873. Papirutgåve i eitt eksemplar oppvaring hos NO 2014
Målførekarta	Digitaliserte bilete	Målførearkivet, UiO	uvurderlege, teikna i perioden 1950-1980	på nettet	Målføre	596 handteikna kart med målføreisoglossar, ca 50 av desse digitaliserte	Papireksemplar finst i eitt eks ved ILN. Dei 50 digitaliserte karta skal registrerast i enkel database esom kopålar saman kart og bakgrunnsinformasjon pr kart, og denne databasen skal planleggast i 2008
lydbandopptak frå Målførearkivet	Lydfiler	Målførearkivet, UiO	uvurderleg originaltilfang på lydband, digitalisert	krev passord, finst på www.dokpro.uio.no/talemaal/	målføre, tale, under transkripsjon i samarbeid med Tekstlaboratoriet	1362 timar, 16,25 Gb data komprimert til 1/10 av opph storleik	registrant over digitaliserte lydbandfiler frå Målførearkivet Sjå sluttrapport_lydband_2005, jfr opplysning frå Tekslaboratoriet
Skogen-samlinga	MSWord-filer	Anders Skogen	stor og gjennomarbeid samling i ms	For NO-redaksjonen. Under arbeid, k-å finst som word-filer	målføre og nynorsk	ca 5000 ordartiklar i alt, ms på	Anders Skogen arbeidde i si tid i NO. Originalmanuskript under digitalisering, vel halvparten innskrivi.

Navn	Ressurstype	Leverandør	Verdi	Tilgang	Målform	Omfang	Kommentar
Haugen-samlinga	Database med digitaliserte bilete	Einar Haugen	stor og gjennomarbeid samling	Under arbeid	Målføre og nynorsk og engelsk		Opphavleg arkivkort-samling frå feltarbeid i Oppdal ved Einar Haugen
Norsk landbruksordbok	Digitalisert (eige format, ved Hjulstad)	Landbruksmållaget	stor og gjennomarbeidd ordbok			ca 35 000 ordartiklar	Skal konverterast til xml og gjerast søkbar via Ordbokshotellet som ordbok.
Norsk Allkunnebok	MSWord-filer	Fonna Forlag og artikkelforfattarane	leksikon på 10 band	For NO-redaksjonen. Under arbeid, eitt band finst som word-filer		10436 spalter på papir, 1 band (960 spalter) digitalisert	skal gjerast søkbart som leksikon, og leggjast inn i korpus i fulltekstutgåve.
Skards ordliste	databaser	Matthias Skard	register over 1938-rettskrivinga	på nettet		35000 ordartiklar	Den mest omfattande ordlista på grunnlag av 1938-rettskrivinga
Ordsamlingar 1600 - 1900	Xml-filer	dei fleste er utgjevne på 1900-talet	digitalutgåvene står for ? tal årsverk i dokproperioden	dei normerte har funnest på nettet, skal opp att	målføre, nynorsk	35 ordsamlingar, svært varierande storleik.	Eldre ordsamlingar, som er med i grunnlaget for Norsk Ordbok. 6 er normerte.
Torp: nynorsk etymologisk ordbok	Xml-tekst	ILN	2 årsverk	Skal koma på nettet i 2008	Nynorsk (1917)	Ca 40 000 ordartiklar	Viktig leksikografisk og etymologisk referansekjelde
Ross: Norsk Ordbog med tillegg 1-6	Word-filer	ILN	? årsverk		Nynorsk (1873)		Viktig leksikografisk referansekjelde. Inkorporert i Grunnmanuskriptet.
Aasen: Norsk Ordbog (1873)		ILN	? årsverk		Nynorsk (1873)		Viktig leksikografisk referansekjelde. Inkorporert i Grunnmanuskriptet.
Aasen: Ordbog over det norske folkesprog (1850)		ILN	? årsverk		Nynorsk, målføre		Viktig leksikografisk referansekjelde. Inkorporert i Grunnmanuskriptet.
Litteraere bokmålstekster	Tekstarkiv over digitaliserte bokmåls-forfatterskap	Forfattere		På nettet	Bokmål	46840 sider tekst	Tekster fra 1550 til 1900
Wittgensteinarkivet	Tekstarkiv	Aksis					Filosofisk tekstarkiv basert på Wittgensteins håndskrevne manuskript (Nachlass).Tekstkritisk xml-koding
Medieval Nordic Text Archive	Tekstarkiv	UiB					Tekstkritisk xml-koding
Henrik Ibsens skrifter	Tekstarkiv	Aksis					Tekstkritisk xml-koding
MPF – Pergament-fragmenter	Tekstarkiv	UiB					Tekstkritisk xml-koding
Eldre tekstmateriale	Tekstkorpus	Aksis					Skjønnlitterarur, Inngår delvis i Oslo-korpuset

Vedlegg C

SVOT-analyse for Språkbanken

Arbeidsgruppa har diskutert sterke og svake sider ved den norske språkbanksatsinga. Me har brukt ein såkalla SVOT-analyse, der me i stikkordsform ser på styrke og veikskap ved modellen vår. Ut frå vurderingane her ser me på kva nye mulegheiter (opningar) satsinga gjev, og kva som trugar ein norsk språkbank. Arbeidsgruppa meiner analysen kan vera eitt av utgangspunkta for dei som skal driva arbeidet med Språkbanken vidare.

Styrke	**V**eikskap	**O**pningar	**T**ruslar
- Dette er ei statleg satsing, som gjev ei viss sikkerheit og kvalitet i vidare arbeid - Språkressursane frå Nordisk Språkteknologi er sikra for språkbanken - Språkrådet har i snart 10 år hatt eit etablert språksekretariat og har arbeidt med språkteknologi i denne perioden - Det finst ein norsk høgkompetanse på feltet. KUNSTI-programmet til NFR har mellom anna sikra solid forsking på feltet - Det er fleire ulike norske forskingsmiljø som har til dels ulik fagleg fordjuping	- Det er ikkje noka fast organisering av fagområdet i Noreg, og modellen vår føreset ein liten organisasjon - Norsk er eit lite språksamfunn og ein liten språkmarknad - Norsk er oppdelt i to skriftmål og ei rad dialektar, noko som aukar omfanget på arbeidet - Det at det er fleire tunge fagmiljø kan også gjera det vanskelegare å få til ei nasjonal organisering - Dei norske språkressursane er spreidde og til dels lite kompatible	- Språkbanken kan sikra betre og meir avansert forsking på feltet - Lett tilgang på språkressursar vil gje vekst for it-næringane og sikra fleire teknologitilbod på norsk - Me kjem nærare målet om norsk som eit komplett samfunnsberande språk - Ei satsing på språkbanken kan gje opning for eit større nordisk samarbeid på feltet, der Noreg kan spela ei leiande rolle.	- Språkbanken får ikkje til samordninga av miljøa, slik at det blir indre konfliktar mellom fagmiljø eller med næringslivet - Usikkerheit omkring storleik på og regularitet av statlege løyvingar - Språkbanken blir lite interessant for mål-gruppene, investeringa gjev manglande meirverdi - Språkbanken skaper nytt byråkrati som ikkje lettar tilgangen på ressursar - At engelsk trass i norsk satsing blir det rådande språket når ny språk-teknologi skal forskast på eller brukast i marknaden

Vedlegg D

VEDTEKTER FOR NORSK SPRÅKBANK

§1
Selskapets navn er Norsk språkbank AS. Til daglig benyttes betegnelsen Språkbanken. I internasjonale sammenhenger benyttes det engelske navnet The Norwegian Human Language Technology Resource Collection med kortformen The Norwegian (HLT) Language Bank.

§2
Selskapets forretningskontor er i ....

§3
Språkbankens formål er å være det sentrale samlingspunktet for lagring og distribuering av offentlige og private digitale språkressurser. Språkbanken er infrastruktur for språkteknologisk forsking, utvikling og produkt-/tjenestetilpassing for norsk språk.

§4
Selskapets aksjekapital er kroner 3 000 000, fordelt på 3 000 aksjer à kr 1 000.

§5
Aksjene eies av staten ved Det kongelige kultur- og kirkedepartementet.

§6
Selskapets styre skal ha sju aksjonærvalgte medlemmer og tre varamedlemmer. Daglig leder er styrets sekretær og møter på styremøtene. Generalforsamlingen velger styreleder.

§7
Selskapets firma tegnes av styrets leder og daglig leder i fellesskap. Styret kan meddele prokura.

§8
Ordinær generalforsamling holdes hvert år innen utgangen av første halvår i selskapets lokaler i Oslo eller etter departementets avgjørelse. Innkalling skjer med åtte dagers skriftlig varsel.

§9
Den ordinære generalforsamlingen skal behandle:

- Fastsettelse av resultatregnskap og balanse
- Anvendelse av overskuddet eller dekning av underskuddet i henhold til den fastsatte balansen, samt utdeling av eventuelt utbytte
- Valg av styre og styrets leder
- Andre saker som i henhold til norsk lov hører inn under generalforsamlingen

Godkjent av generalforsamlingen xx.xx.200x.

Vedlegg E

ARBEIDSLISTE – SPRÅKBANKEN

Mulig forarbeid høsten 2008:

- detaljert kartlegging av identifiserte språkressurser:
  - hva finnes hvor
  - avklare opphavsrettslige problemstillinger
  - vurdere kvaliteten
  - anslå utgiftene til bearbeiding og tilrettelegging for Språkbanken
- grundig gjennomgang av BLARK (gir bedre grunnlag for prioritering)
- utarbeide kontraktsformularer for avlevering og distribusjon av bruksretten til språkressurser til/fra Språkbanken
- utarbeide samtykkeformular som kan brukes ved innsamling av talemateriale

Etablering:

1. Stortinget vedtar å opprette Språkbanken fra 1. januar 2009.
2. Stortinget vedtar budsjettet for Språkbanken (statsbudsjettet).
3. Kultur- og kirkedepartementet vedtar vedtekter for Språkbanken.
4. Kultur- og kirkedepartementet oppnevner styret for Språkbanken.
5. Styret lyser ut stillingen som daglig leder for Språkbanken.
6. Daglig leder sørger for nødvendige dokumenter til Brønnøysundregistrene.
7. Daglig leder (sammen med styret) lyser ut og tilsetter i de to andre stillingene.
8. Administasjonen etableres, kontorene utstyres, administrative rutiner lages.
9. Retningslinjer for avleverings- og distribusjonsformater og –rutiner spesifiseres.
10. Innlemming av eksisterende ressurser kan starte, nødvendig nyinnsamling (komplettering av eksisterende ressurser) kan settes i gang.
11. Nettsidene oppdateres med relevant informasjon (finnes - må overføres, domenenavn er kjøpt – må overføres) www.sprakbanken.uib.no må bli til www.sprakbanken.no