Gjennomgang og evaluering av språkressurser fra NSTs konkursbo

Rapport

Utarbeidet av

Gisle Andersen
Forsker/prosjektkoordinator

Avdeling for kultur, språk og informasjonsteknologi (Aksis) ved
UNIFOB/Universitetet i Bergen

Desember 2005

Innhold

Gjennomgang og evaluering av språkressurser fra NSTs konkursbo

1 Innledning

Dette dokumentet inneholder en gjennomgang og evaluering av NSTs språkressurser. Det innledende avsnittet gir bakgrunn for evalueringsarbeidet og beskriver metode. Den første hoveddelen, avsnitt 2, inneholder en gjennomgang av de mest relevante delene av NSTs maskinpark. I den andre hoveddelen, avsnitt 3, finnes beskrivelsen av selve språkressursene. I avsnitt 4 gis en kort sammenfatning og anbefalinger knyttet til ressursene.

1.1 Bakgrunn

Selskapet Nordisk Språkteknologi Holding AS gikk konkurs i november 2003. På det tidspunkt hadde selskapet gjennom en årrekke bygget opp svært omfattende språkressurser. Språkrådet har tatt initiativ til å fremskaffe mer detaljert informasjon om språkressursene enn det som hittil har vært tilgjengelig, noe NSTs konkursbo ved bostyrer Arne Laastad stilte seg positiv til. Til å utføre oppdraget kontaktet Språkrådet derfor Avdeling for kultur, språk og informasjonsteknologi (Aksis) ved UNIFOB/ Universitetet i Bergen, som har påtatt seg oppdraget.

Undersøkelsen har hjemmel i avtale inngått 12. juli 2005 mellom Språkrådet og Nordiske Språkressursar AS/Nordisk Språkteknologi Holding AS' konkursbo.

Undersøkelsen er koordinert og i all hovedsak foretatt av forsker/prosjektkoordinator Gisle Andersen ved Aksis. Konsulent Tore Burheim og IT-ansvarlig Lars Rørlien har bistått med infrastruktur, opplysninger og tilgang til NSTs tidligere lokaler og maskinpark. Språkrådets rådgiver Torbjørg Breivik har bidratt med vurdering av avtaleverk knyttet til språkressursene. Vurderingen foreligger i en egen rapport.

1.2 Undersøkelsens omfang og innhold: språkressurser

Med språkressurser menes i denne sammenheng *språklige primærressurser* som er egnet som grunnlag for å produsere språkteknologisk programvare. Dette omfatter i hovedsak tre typer primærressurser: tekstdatabaser (korpus), taledatabaser og leksikalske databaser. I mindre grad er det som kan kalles sekundærressurser omhandlet. Disse omfatter språkbehandlingsverktøy og språkteknologisk programvare som er utviklet på grunnlag av de beskrevne primærressursene.

Gjennomgangen har bestått i å kartlegge og beskrive aktuelle arkiver som kan tenkes å inneholde språkressurser. Dette dreier seg om flere ulike lagringsformer: servere, personlige datamaskiner, CD-arkiv og tape. Gitt den tilgjengelige tidsrammen er den detaljerte undersøkelsen begrenset til serverne og CD-arkivet. Tapeformatet er kun brukt for tidligere sikkerhetskopiering; tapene er for øvrig vanskelig tilgjengelig, og det er ikke ansett som hensiktsmessig å gjennomgå dem som del av gjeldende undersøkelse. En del av NSTs maskinpark har blitt solgt. Dette omfatter imidlertid hovedsakelig kun personlige datamaskiner. For de PC-er som har blitt solgt er det gjort fullkopi i form av en såkalt ghost. En gjennomgang av personlige datamaskiner er ikke foretatt. Gjennomgangen av CD-arkivet er stikkprøvebasert. Det forventes ikke at lagringsformater som ikke er gjennomgått inneholder språkressurser av

betyding, men det antas at de kan inneholde kopier av materialet som er gjennomgått. Denne rapporten tar dermed mål av seg å være *fullstendig* med hensyn til kartlegging av NSTs språklige primærressurser.

Denne vurderingen omhandler også en eventuell vanskeliggjøring av tilgjengelighet eller forringelse av data som selve konkursen – og det faktum at ressursene har stått uberørt i en lang stund – måtte ha medført.

1.3 Metode for kartlegging og beskrivelse av ressursene

Gjennomgangen av ressursene er foretatt med det formål å dokumentere

- ressursenes plassering på serverne
- format
- størrelse
- kvalitet.

Vurdering av det siste punktet er foretatt med utgangspunkt i gjeldende dokumentasjon fra organisasjonen ELRA (European Language Resources Association).[1]

For leksikalske ressurser er gjennomgangen fullstendig i den forstand at alle leksikon har vært inspisert under gjennomgangen, mens det kun er foretatt stikkprøvekontroll av enkeltposter i leksikonene. De akustiske ressursene er vurdert ved stikkprøvekontroll av lydfiler annoteringsfiler og manuskriptfiler. Korpusressursene har blitt gjennomgått ved kontroll av enkelttekster fra hver enkelt leverandør.

Gjennomgangen er foretatt ved å undersøke én server av gangen, og beskrive de kataloger som inneholder språklige primærressurser. De kataloger som inneholder annet materiale (f.eks. programvare eller korrespondanse) er kun beskrevet i den grad de fremstår som særlig relevant.

2 Systematisk gjennomgang av servere

Gitt NST-ressursenes store omfang synes det riktig å starte gjennomgangen med en kartlegging av det som finnes i NSTs maskinpark. I dette avsnittet finnes en kort beskrivelse av hver enkelt server og en omtale av dens funksjon og innhold og hvor relevant serveren er med hensyn til språkressurser. Hensikten med denne beskrivelsen er å gjøre det enklere å finne frem i materialet ved senere bruk.

Avsnitt 2.1 gir en oversikt i form av tabulære skjema over serverne som finnes på Voss.[2] Det er kun foretatt detaljert gjennomgang av filer og kataloger som kan tenkes å inneholde språkressurser. Disse er beskrevet i detalj i påfølgende avsnitt.

---

[1] Dette gjelder i hovedsak følgende dokumenter: http://www.spex.nl/validationcentre/d11v21.doc (akustisk database), http://www.elra.info/services/validation_manual_lexica.pdf (leksikon), og http://www.elra.info/services/valid/wp3/index.htm (korpus).

[2] Det er brukt hyperlenker i den elektroniske versjon av denne teksten. For at disse skal fungere må leseren være logget på NSTs nettverk på Voss og ha den nødvendige servertilgang.

2.1 Oversikt

2.1.1 Windows-servere

Serverne befinner seg fysisk på NSTs serverrom på Tvildemoen. Serverne \\Main og \\Server2 anses som de viktigste Windows-serverne med hensyn til språkressurser og er gjennomgått i mer detalj i påfølgende avsnitt.

Windows-server	Stasjon	Funksjon/innhold	Gjennomgått	Inneholder språkressurser
Main	\\Main	Ressursoppbygging	Ja	Ja
Server1	\\Server1	Systemutvikling	Delvis	Kun sek.ress
Server2	\\Server2	Ressursoppbygging	Ja	Ja
Dilbert	\\Dilbert	E-postserver	Nei	Nei
Rasserver	\\Rasserver	Kommunikasjonsserver	Nei	Nei

2.1.2 Linux-servere

De seks Linux-serverne ble kjøpt inn i forbindelse med inngått avtale om felles programvareutvikling mellom NST og IBM og befinner seg nå på NSTs serverrom på Tvildemoen. Samtlige servere er beskrevet i mer detalj under gjennomgangen av språkressurser i avsnitt 3.

Linux-server	Stasjon	Funksjon/innhold	Gjennomgått	Inneholder språkressurser
Rene	\\Rene	TTS-utvikling norsk og svensk	Ja	Ja
Heidi	\\Heidi	TTS-utvikling dansk	Ja	Ja
Michelle	\\Michelle	Akustisk modellering norsk	Ja	Ja
Kate	\\Kate	Akustisk modellering svensk	Ja	Ja
Ursula	\\Ursula	Akustisk modellering dansk	Ja	Ja
Mary	\\Mary	Diktering norsk og svensk	Ja	Ja

2.1.3 CD-arkiv

NSTs CD-arkiv utgjør ca. 70 arkivbokser med CD-er som inneholder originaler og sikkerhetskopier av språkdata, i tillegg til øvrig sikkerhetskopiering av dokumentasjon, programvare etc. Både arkivbokser, CD-futteraler og CD-plater er merket med innhold. Dette har vært gjennomgått etter stikkprøveprinsippet. CD-arkivet er en viktig ressurs fordi det inneholder akustiske data som ikke later til å være representert i noen annet lagringsformat (jf. 3.1).

2.1.4 Personlige datamaskiner

En del av NSTs maskinpark har blitt solgt. Dette omfatter så godt som samtlige brukermaskiner som stod plassert på kontorer, og noen mindre servere. For maskiner som ikke lenger er en del av boet har man tatt vare på alt innhold i form av en fullkopi (et såkalt Ghost Image). Denne er produsert ved hjelp av programmet Norton Ghost. Dette materialet finnes på serveren \\Nas-voss. Her finnes 59 filer som omfatter 396 GB data. Disse er av formatet **01010.gho** og **010101.ghs**, hvor filnavnet indikerer hvilken maskin det er snakk om. Filekstensjonen ***.gho** er hovedfilen for enkeltmaskiner, mens ***.ghs** er fortsettelse av kopiering av den enkelte maskin.

Det finnes også fullkopier av 24 datamaskiner som befinner seg på \\Server2\Image.

Dette materialet har ikke vært gjennomgått som en del av undersøkelsen. Det antas at det ikke inneholder språklige grunnressurser som ikke allerede finnes på en av de gjennomgåtte serverne eller i CD-arkivet. Tilgang til materialet krever installering av Norton Ghost.

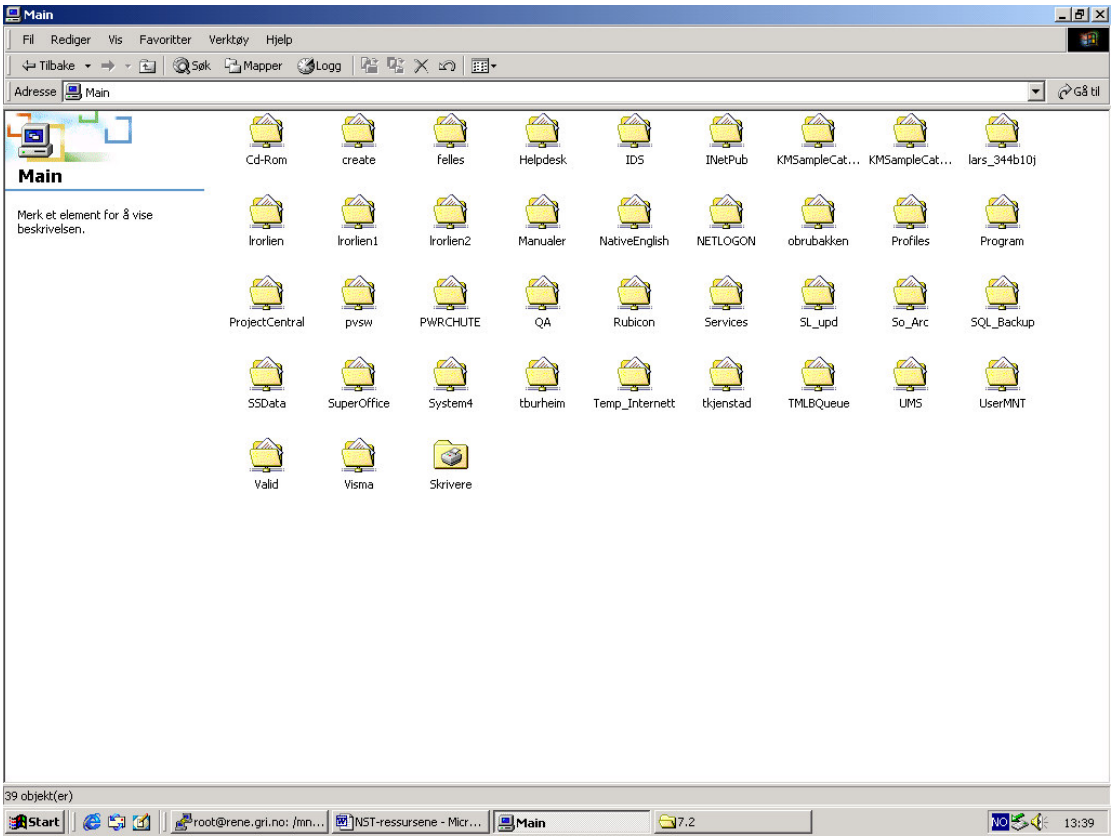
2.1.5 Streamertape

Et uvisst antall såkalte DLT streamertapes befinner seg i NSTs safe og i rommet med CD-arkivet. Ressursene er vanskeligere tilgjengelig og har ikke vært gjennomgått som en del av undersøkelsen. Lagringsformatet er benyttet til sikkerhetskopiering av data som også befinner seg på servere. Dette er beskrevet mer utførlig under hver enkelt ressurs.

2.2 Gjennomgang av serveren Main

Foruten språkressurser inneholder serveren **Main** meget annet innhold. Dette omfatter blant annet programvare, fullversjoner av programmanualer, dokumentasjon og personlige dokumenter.

Serveren inneholder kataloger som vist i figuren.



Kun et subsett av disse har vært tilgjengelig for undersøkelsen, men det er på det rene at det som fins av språkressurser har vært gjennomgått. Tabellen viser kataloger som inneholder aktuelle språkressurser:

Katalog	Innhold	Relevans
\\Main\Felles	Her finnes store deler av språkressursene. Gjennomgått i detalj og beskrevet nedenunder.	Språkressurser
\\Main\InetPub	Microsoft Index Server	Ikke språkressurser
\\Main\Manualer	Et 90-talls bøker, hovedsaklig brukermanualer for programvare	Ikke språkressurser
\\Main\NativeEnglish	Programvare og dokumentasjon for L&H-programmet Correct English, skandinaviske versjoner utviklet NST, regelsett, grammatikker m brukerkommentar utviklet for da/no/sv.	Sekundærressurser
\\Main\NETLOGON		Ikke språkressurser
\\Main\Program	Et bredt utvalg programvare	Ikke språkressurser
\\Main\ProjectCentral	Microsoft Project Central	Ikke språkressurser
\\Main\Pvsw		Ikke språkressurser
\\Main\Services	SLServer	Ikke språkressurser
\\Main\SL_upd		Ikke språkressurser
\\Main\So_Arc	Administrativt	Ikke språkressurser
\\Main\SSData	Administrativt	Ikke språkressurser
\\Main\SuperOffice	Administrativt	Ikke språkressurser
\\Main\Temp_Internett		Ikke språkressurser
\\Main\UserMNT		Ikke språkressurser
\\Main\Valid	Her finnes dokumentasjon knyttet til NSTs akustiske database, og noen primærressurser	Språkressurser
\\Main\Skrivere		Ikke språkressurser

2.2.1 Oversikt over området Felles på Main

Partisjonen \\Main\Felles er gjennomgått i detalj. Den inneholder språklige primærressurser, verktøy og administrativ informasjon. Det er særlig leksikalske ressurser som finnes her, som vist i tabellen.

Katalog	Innhold	Relevans
\\Main\Felles\Assistive technologies and learning systems	Hovedsakelig forretningsmessig dokumentasjon, men også programvare, særlig TTS-produktene ReadIt og Small Talk	Sekundærressurser: programvare
\\Main\felles\Holding	Administrativt	Ikke språkressurser
\\Main\felles\HowTo	Administrativt	Ikke språkressurser
\\Main\felles\Konsern	Administrativt	Ikke språkressurser
\\Main\felles\Nst	Administrativt	Ikke språkressurser
\\Main\felles\Oslo	Administrativt	Ikke språkressurser
\\Main\felles\TIT	Språkressurser, leksikon, verktøy, dokumentasjon osv. omtalt i detalj under beskrivelsen av enkeltressurser	Språkressurser
\\Main\felles\TIT\Admin	Dokumentasjon knyttet til ressursutvikling, prosjektrapporter	Sekundærressurser: dokumentasjon
\\Main\felles\TIT\Diverse	Dokumentasjon knyttet til ressursutvikling, rapporter, verktøy, bl.a. ..\Diverse\Links_og_rapporter\ Generelle rapporter\Generelle rapporter\LDBTools	Sekundærressurser: dokumentasjon
\\Main\felles\TIT\Projekter	Hvor selve arbeidet med språkressurser har foregått, særlig i IBM-tiden	Primær- og sekundærressurser

\\Main\felles\TIT\Ressourcer	Akustiske ressurser, leksikalske ressurser, korpus. Beskrevet i detalj senere.	Primærressurser

2.2.2 Oversikt over området Valid på Main

Partisjonen \\Main\Valid er gjennomgått i detalj. Den inneholder administrativ informasjon knyttet til NSTs akustiske database, men ikke selve dataene.

Katalog	Innhold	Relevans
\\Main\Valid\ADB_OD	Administrativt	Ikke språkressurser
\\Main\Valid\ADB_indspilninger_mv	Database over informanter	Dokumentasjon
\\Main\Valid\ADB_Telefoni	Arbeidsområde for telefonidata	Dokumentasjon
\\Main\Valid\opptak_danmark	Administrativt	Ikke språkressurser
\\Main\Valid\opptak_norge	tom	-
\\Main\Valid\PERL_TOOLS	Verktøy for utregning av informantstatistikk	Sekundærressurs

2.3 Gjennomgang av Server1

Server1 ble hovedsakelig brukt av NSTs systemutviklergruppe og inneholder programvare, verktøy og dokumentasjon knyttet til systemutvikling. Det antas at det ikke er primærressurser på serveren. Dette har imidlertid ikke latt seg kontrollere ordentlig, fordi det meste av serveren ikke var tilgjengelig for inspeksjon.

2.4 Gjennomgang av Server2

Server2 inneholder hovedsakelig språkressurser, i særdeleshet tekstkorpus. Den overordnete filstrukturen er som følger:

Katalog	Innhold	Relevans
\\Server2\ABLEX	Data, verktøy og dokumentasjon av NFR-prosjektet Arbeidsbenk for leksikon og korpus (KUNSTI-programmet)	Primærressurser: dataverktøy
\\Server2\Image	Ghost images fra en del datamaskiner,	Ikke språkressurser
\\Server2\KorpusOrginal	Originalversjoner av dansk, norsk og svensk tekstkorpus	Primærressurser: korpus
\\Server2\KorpusWork	Bearbeidete versjoner av dansk, norsk og svensk tekstkorpus	Primærressurser: korpus
\\Server2\OFA	Loggfiler over leveranser til svensk Office diktering.	Ikke språkressurser
\\Server2\OFFICE	Programvare	Ikke språkressurser
\\Server2\Skrivere	Tom	Ikke språkressurser

3 Språkressurser

Beskrivelsen av NST-ressursene i dette avsnittet er organisert etter ressurstype og omfatter akustiske databaser for talegjenkjenning (avsnitt 3.1), akustiske databaser for talesyntese (avsnitt 3.2), leksikalske databaser (avsnitt 3.3) og korpus (avsnitt 3.4).

3.1 Akustiske databaser for talegjenkjenning

De akustiske databasene omfatter to hovedtyper: databaser produsert for talegjenkjenning/diktering, og databaser produsert for talesyntese. Førstnevnte kategori er den desidert mest omfattende. Akustiske databaser for talesyntese er beskrevet i avsnitt 3.2.

I dette avsnittet følger først, i avsnitt 3.1.1, en generell beskrivelse av NST database for talegjenkjenning/diktering. Deretter følger språkspesifikke ressursoversikter for norsk, svensk og dansk i avsnittene 3.1.2-3.1.4, med beskrivelser av ressursenes omfang, plassering, formater og valideringsgrad. Avsnitt 3.1.5 gir en beskrivelse av den metode og standard som ligger til grunn for validering av opptakene. Avsnitt 3.1.6-3.1.7 tar for seg spesifikke språklige tema og dialektområder. Deretter følger en samlet kvalitativ vurdering av de akustiske ressursene i avsnitt 3.1.8.

3.1.1 Generelt om NSTs akustiske database

NSTs akustiske database ligger i sin helhet i CD-arkivet som befinner seg på Voss. En betydelig del av materialet har blitt overført til servere i forbindelse med produksjon av akustiske modeller og leveranser til L&H/IBM. Det er foretatt sikkerhetskopiering av det meste av dataene i form av DLT streamertape som befinner seg i NSTs safe. Disse har ikke vært en del av undersøkelsen.

Materialet fordeler seg på ulike kategorier definert av formålet for innspillingen. Disse er beskrevet nedenfor. Tabellene i påfølgende språkspesifikke avsnitt gir oversikt over hvilke data som er lagret på streamertape og servere.

Generelt for databasen gjelder at den er i sin helhet samlet inn og validert av NST selv. Unntak fra dette er deldatabasene SpeechDat og Telia, som er innkjøpte baser og del av såkalte ”in kind”-ressurser som NST anskaffet ved aksjeemisjon. SpeechDat består av data for mobiltelefoni og fasttelefoni for norsk, svensk og dansk. NST har ikke eierrettighetene til dette materialet, og det er således ikke omtalt i påfølgende avsnitt. Denne ressursen er godt dokumentert andre steder (jf. http://www.speechdat.org/ og http://www.elda.org/catalogue_en.php?speechdat&and) Telia-materialet består av kontormiljøbaserte innspillinger for talegjenkjenning gjort i Stockholmsområdet.

Hoveddelen av de akustiske ressursene er spilt inn og validert ved hjelp av programvare fra L&H. Opptaksprogrammet DSDR (Desktop Speech Digital Recorder) er benyttet, så sant ikke noe annet er opplyst nedenfor. All validering har foregått ved hjelp av valideringsprogrammet DSVS (Desktop Speech Validation Station). Bruken av programvaren er beskrevet i følgende dokumenter: \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_T\Norsk\Validering\Sjekkliste Telefoni.doc og \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_OD\Validering\VALIDATION DICTATION-ASR-V2.doc.

Tilgang til denne proprietære programvaren er imidlertid ingen forutsetning for å nyttiggjøre seg dataene til fremtidige formål. Selve dataene foreligger nemlig i generelt anvendelige formater, i form av lydfiler i PCM/wav-format og spl-loggfiler i

rent tekstformat. (Mer utførlig beskrivelse følger.) Lydfilene ligger lagret i ukomprimert form (ikke bruk av zip eller tilsvarende verktøy).

Katalogen \\Main\felles\TIT\Projekter\Resursorg\TIT0022\Docs inneholder dokumentasjon av NSTs akustiske database. En ressursoversikt finnes også i dokumentet \\Main\felles\TIT\Projekter\Resursorg\TIT0022\adb.xls. I det følgende gis en kvantitativ og kvalitativ beskrivelse av databasens innhold.

3.1.2 Norsk

3.1.2.1 Opptak for talegjenkjenning (ASR/Diktering), 16 kHz

Deldatabasen **ADB_OD_Nor.NOR** er samlet inn for produksjon av teknologi for akustisk modellering for PC/Multimedia talegjenkjenning og for automatisk diktering (Office ASR and Dictation). Opptakene er gjort i lukket kontormiljø og baserer seg på fonetisk balanserte manuskript, produsert på grunnlag av setninger fra NSTs norske korpus. Databasen består av en treningsdel og en testdel, der førstnevnte brukes til å trene selve den akustiske modellen, mens sistnevnte brukes for testformål. Fordelingen av opptak for de to delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Trening	nor0463	312	900	280800	253 CD-er + tape	97,5	Mary
Testing	nor0464	987	80	78960	81 CD-er + tape	26,9	-

Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innpilt enhet, dvs. som oftest en setning eller noen tilfeller en frase eller enkeltord. Deldatabasen finnes ytterligere dokumentert på \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_OD. Databasen slik den foreligger på CD er organisert etter en bestemt katalogstruktur, som vist nedenfor.

Lydfiler	D:\adb_0464\speech\scr0464\23\04642301\r4640007
Annoteringsfil	D:\adb_0464\data\scr0464\23\04642301**r4640007.spl**
Liste over annoteringsfiler	D:\adb_0464\doc**Spl.lst**
Instruksjoner til informant	D:\adb_0464\doc**nor464.scr**

Konvensjonene for navngiving av spl-filene i katalogen **data** er følgende: …/skriptnummer/stasjonsnummer/gruppenummer/loggfil. Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	16 kHz
Oppløsning	16 bit
Format	Intel PCM
Kanaler	2 (stereo)

Dette formatet er i overensstemmelse med kravspesifikasjonene fra L&H. En del av materialet er konvertert fra dette utgangspunktet til et format som kreves for produksjon av akustiske modeller basert på IBM-teknologi. Disse opptakene finnes på følgende server:

\\Mary\felles_mary\ac1\data\dyf (kvinner)
\\Mary\felles_mary\ac1\data\dym (menn)

Denne deldatabasen består av de 312 treningsopptakene fra 415 menn og 485 kvinner. For dette materialet gjelder følgende teknisk informasjon:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	16 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	1 (mono)

For hver informant ligger lydfilene og manuskriptfilene henholdsvis på
\\Mary\felles_mary\ac1\data\dym\a03\no\16.pcm
\\Mary\felles_mary\ac1\data\dym\a03\no\scripts\Word

Innspillingsskriptet som treningsdataene bygger på har en dikteringsdel og en ASR-del. Førstnevnte del består av ordinære korpusekstraherte setninger som kreves til generelle dikteringsformål, og utgjør manuskriptets 222 første enheter (setninger). Sistnevnte del omfatter de siste 90 enhetene og består av fraser som utgjør personnavn, stedsnavn, enkeltord, akronymer og annet som kreves til spesifikke talegjenkjenningsformål (ASR). Tegnsetting er eksplisitt lest.

Innspillingsskriptet som testdataene bygger på har en tilsvarende inndeling i en dikteringsdel og en ASR-del. Dikteringsdelen utgjør manuskriptets første 750 enheter, mens ASR-delen utgjør de siste 237 enhetene.

Hele materialet har blitt validert etter de kriterier og metoder som er nevnt i avsnitt 3.1.5. Valideringsfilene ligger på \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Filer\Norsk\ADB_OD_Nor.NOR\Validering\.

3.1.2.2 Opptak for diktering, 22 kHz

Deldatabasen **ADB_D_IBM-N** er samlet inn for produksjon av teknologi for akustisk modellering for automatisk diktering (desktop). Ulikt foregående ressurs er opptakene spilt inn ved hjelp av IBM-programvaren ObjectRexx.[3] Opptakene ble gjort i forbindelse med oppstart av samarbeidet mellom NST og IBM som ledd i opplæringsperioden av NST-ansatte. Databasen består av tre deler innspilt til ulike formål: en testdel, en treningsdel og en modelleringsdel. Fordelingen av opptak i delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Modellering	mod	260	104	27040	83 CD-er	6,24	-
Testing	test	160	20	3200		0,9	-
Enrollment	enroll	156	20	3120		0,9	-

Som oversikten viser så er disse dataene ikke funnet på andre lagringsformater enn i CD-arkivet. Følgende teknisk informasjon gjelder for denne deldatabasen:

[3] \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_D_IBM\Recordings\IBM_object_rexx_program.

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	22 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	1 (mono)

Skriptene som opptakene er basert på ligger også på CD og har filnavn **no_mod_001.txt** (002, 003 etc), **no_enroll.txt** og **no_test.txt**. Disse finnes også på \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_D_IBM\Recordings\Innspillingsskript\Norsk\Klart_no. Opptakene er gjort i lukket kontormiljø, og baserer seg på fonetisk balanserte manuskript, produsert på grunnlag av avistekst fra Aftenpostens 1996-årgang. Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innspilt enhet (setning, frase, enkeltord, tallrekke, bokstavrekke).

Denne deldatabasen er ikke validert. Det foreligger derfor begrenset dokumentasjon, men en del informasjon finnes i dokumentet \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_D_IBM\Team_organisering\sluttrapportADB_IBM.doc. Mikrofonen som ble benyttet er en Andrea NC-61, og lydkortet er Turtle Beach Montego II.

3.1.2.3 Opptak for telefoni, 8 kHz

Deldatabasen **ADB_T_Nor.NOR** inneholder opptak til telefoni, fordelt på fastlinje og mobiltelefoni. Disse er egnet til bruk ved produksjon av takegjenkjenningsteknologi for telefoni. Materialet er ikke inndelt i test- og treningsdata. NST har fulgt de generelle SpeechDat II-prosedyrene under opptak. Opptakene er gjort delvis med L&Hs programvare og delvis ved UMS Diginform.

Informantene har fått oppgitt telefonnummer og ringt inn og lest opp setninger. Opptakene inneholder 17 ytringer som er spontan tale i form av svar på spørsmål, og 40 ytringer med oppleste manuskriptsetninger. Skriptet **nor0531.scr** ble brukt både til fasttelefoni og mobiltelefoni. I forbindelse med utvikling av en ASR-applikasjon for NSB ble det også gjort opptak av navn på norske jernbanestasjoner. Disse er omfattet av manuskriptet **nor0666.scr**. Tabellen viser fordelingen av disse opptakene.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str GB	Server
Fasttelefoni	nor0531.scr	57	6231	355167	571 CD-er + tape	27,2	Michelle
Mobiltelefoni	nor0531.scr	57	2018	115026			
Fasttelefoni	nor0666.scr	101	65	6565	18 CD-er + tape	0,4	Michelle
Mobiltelefoni	nor0666.scr	101	37	3737			

Deldatabasen finnes i sin helhet i NSTs CD-arkiv. Det er foretatt fullstendig sikkerhetskopi på streamertape som ligger i safen. I tillegg finnes det CD-er og JAZ-tapes med råopptak som NST fikk tilsendt fra under leveranse fra UMS. Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding	mu-Law
Filformat	wav
Samplingsfrekvens	8 kHz
Oppløsning	16 bit
Format	8-bit mu-Law Compressed
Kanaler	1 (mono)

Deldatabasen knyttet til NSB-prosjektet finnes også på serveren Michelle, på \\Michelle\felles_michelle\666\adb_666\speech\scr0666

I dette lagringsformatet foreligger dataene i IBM-kompatibelt telefoniformat, og følgende spesifikasjon gjelder:

Signalkoding	A-Law
Filformat	wav
Samplingsfrekvens	8 kHz
Oppløsning	16 bit
Format	8-bit A-Law Compressed
Kanaler	1 (mono)

Denne deldatabasen er kun delvis validert. Validering av 3108 fastlinje- og 1596 mobilopptak er foretatt, og valideringsfilene ligger på \\Main\felles\TIT\Ressourcer\ADB_oversigt\Filer\Norsk\ADB_T_Nor.NOR

Under valideringen har en del filer har blitt lagt til side fordi de er av for dårlig kvalitet; disse fins i underkatalogen **…\forkastede opptak**. En del opptak har blitt lagt til side av andre grunner enn dårlig kvalitet, f.eks. for å sikre optimal fordeling av informantene. Disse er ikke validerte og ligger i underkatalogen **…\Opptak på vent**. Ferdig validerte filer ligger i underkatalogen **…\Validering**. I tillegg finnes det cirka 1000 opptak som ikke har vært igjennom valideringsprosessen i det hele tatt.

Hele NSB-delen av databasen er validert, og valideringsfilene ligger på \\Main\felles\TIT\Ressourcer\ADB_oversigt\Filer\Norsk\ADB_T_Nor.NOR\Validering\Prosjekt666.

Deldatabasen er ytterligere dokumentert på \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_T.

3.1.2.4 Database med innspilte nølelyder

En deldatabase ble samlet inn for produksjon av spesifikke modeller for nølelyder, dvs. ikke-verbale lyder som uttales når en taler nøler mellom ord. Nølelydene omfatter en nasal og en vokal transkribert som <mmm> eller <eeeh> i manuskriptene. Et slikt materiale er et tillegg som brukes ved produksjon av generelle dikteringssystemer. Materialet ble spilt inn sammen med databasen **ADB_OD_Nor.NOR** beskrevet i avsnitt 3.1.2.1. De samme tekniske spesifikasjonene gjelder for dette subsettet.

Materialet finnes på to CD-er i arkivet; det er ikke funnet i andre lagringsformater. I likhet med resten av ADB_OD_Nor.NOR består det av en treningsdel og en testdel, som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Trening	nor4631	28	10	200	2 CD-er	0,6	-
Testing	nor4641	58	2	100			

Treningsmanuskriptet består av 20 vanlige setninger med to nølelyder i hver setning samt fire isolerte repetisjoner av hver nølelyd. Testmanuskriptet består av 50 vanlige

setninger med to nølelyder i hver setning samt fire isolerte repetisjoner av hver nølelyd.

3.1.3 Svensk

For svensk er det ikke gjort 8 kHz-innspillinger for telefoni; man har kun støttet seg til de innkjøpte SpeechDat-ressursene.

3.1.3.1 Opptak for talegjenkjenning (ASR/Diktering), 16 kHz

Deldatabasen **ADB_OD_Swe.SWE** er samlet inn for produksjon av teknologi for akustisk modellering for PC/Multimedia talegjenkjenning og for automatisk diktering (Office ASR and Dictation). Opptakene er gjort i lukket kontormiljø, og baserer seg på fonetisk balanserte manuskript, produsert på grunnlag av setninger fra NSTs svenske korpus. Databasen består av en treningsdel og en testdel, der førstnevnte brukes til å trene selve den akustiske modellen, mens sistnevnte brukes for testformål. Fordelingen av opptak for de to delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Trening	swe0467	312	920	287040	255 CD-er + tape	96,5	Mary
Testing	swe0468	987	80	78960	74 CD-er + tape	22,7	-

Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innspilt enhet, dvs. som oftest en setning eller noen tilfeller en frase eller enkeltord. Deldatabasen finnes ytterligere dokumentert på \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_OD. Databasen er organisert etter samme katalogstruktur som beskrevet for norsk i avsnitt 3.1.2.1.

Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	16 kHz
Oppløsning	16 bit
Format	Intel PCM
Kanaler	2 (stereo)

Dette formatet er i overensstemmelse med kravspesifikasjonene fra L&H. En del av materialet er konvertert fra dette utgangspunktet til et format som kreves for produksjon av akustiske modeller basert på IBM-teknologi. Disse opptakene finnes på følgende server:

\\Mary\felles_mary\ac1\data\dxf (kvinner)
\\Mary\felles_mary\ac1\data\dxm (menn)

Denne deldatabasen består av de 312 treningsopptakene fra 437 menn og 491 kvinner. For dette materialet gjelder følgende teknisk informasjon:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	16 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	1 (mono)

For hver informant ligger lydfilene og manuskriptfilene henholdsvis på
\\Mary\felles_mary\ac1\data\dxf\a01\sv\16.pcm og
\\Mary\felles_mary\ac1\data\dxf\a01\sv\scripts\Word.

Innspillingsskriptet som treningsdataene bygger på har en dikteringsdel og en ASR-del. Førstnevnte del består av ordinære korpusekserperte setninger som kreves til generelle dikteringsformål, og utgjør manuskriptets 222 første enheter (setninger). Sistnevnte del omfatter de siste 90 enhetene og består av fraser består av fraser som utgjør personnavn, stedsnavn, enkeltord, akronymer og annet som kreves til spesifikke talegjenkjenningsformål (ASR).

Innspillingsskriptet som testdataene bygger på har en tilsvarende inndeling i en dikteringsdel og en ASR-del. Dikteringsdelen utgjør manuskriptets første 750 enheter, mens ASR-delen utgjør de siste 237 enhetene.

Hele materialet har blitt validert etter de kriterier og metoder som er nevnt i avsnitt 3.1.5. Valideringsfilene ligger på
\\Main\Felles\TIT\Ressourcer\ADB_oversigt\Filer\Svensk\ADB_OD_Swe.SWE\Validering\Diktering og
\\Main\Felles\TIT\Ressourcer\ADB_oversigt\Filer\Svensk\ADB_OD_Swe.SWE\Validering\ASR.

3.1.3.2 Opptak for diktering, 22 kHz

Deldatabasen **ADB_D_IBM-S** er samlet inn for produksjon av teknologi for akustisk modellering for automatisk diktering (desktop). Ulikt foregående ressurs er opptakene spilt inn ved hjelp av IBM-programvaren ObjectRexx. Opptakene ble gjort i forbindelse med oppstart av samarbeidet mellom NST og IBM som ledd i opplæringsperioden av NST-ansatte. Databasen består av tre deler innspilt til ulike formål, en testdel, en treningsdel og en modelleringsdel. Fordelingen av opptak i delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Modellering	mod	260	155	40300	83 CD-er	9,3	Mary
Testing	test	160	20	3200		0,9	-
Enrollment	enroll	106	20	2120		0,9	-

Opptakene fins i sin helhet i CD-arkivet. Det er ikke tatt sikkerhetskopi på streamertape av disse. Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	22 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	1 (mono)

Skriptene som opptakene er basert på ligger også på CD og har filnavn **sv_mod_001.txt** (002, 003 etc), **sv_enroll.txt** og **sv_test.txt**. Disse finnes også på \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_D_IBM\Recordings\Innspillingsskript\Svensk. Opptakene er gjort i lukket kontormiljø, og baserer seg

på fonetisk balanserte manuskript, produsert på grunnlag av nyhetstekst fra NSTs svenske korpus. Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innpilt enhet (setning, frase, enkeltord, tallrekke, bokstavrekke).

Som tabellen ovenfor viser, finnes modelleringsdelen av denne deldatabasen på serveren Mary, nærmere bestemt i katalogene

\\Mary\felles_mary\ac1\data\dsf (kvinner)
\\Mary\felles_mary\ac1\data\dsm (menn)

Den samme tekniske spesifikasjonen gjelder for denne kopien av datasettet. Her er lydfilene og manuskriptfilene lagret etter samme struktur som for den norske del av databasen:

\\Mary\felles_mary\ac1\data\dsf\001\sv\22.pcm (lydfiler)
\\Mary\felles_mary\ac1\data\dsf\001\sv\scripts\Word (manuskriptfiler)

Denne deldatabasen er ikke validert. Det foreligger derfor begrenset dokumentasjon, men en del informasjon finnes på
\\Main\Felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_D_IBM

3.1.3.3 Database med innspilte nølelyder

En deldatabase ble samlet inn for å lage spesifikke modeller for nølelyder, på samme måte som for norsk. Nølelydene omfatter to vokaler og en nasal, transkribert som <öööh>, <aaah> eller <mmm> i manuskriptene. Et slikt materiale er et tillegg som brukes ved produksjon av generelle dikteringssystemer. Materialet ble spilt inn sammen med databasen **ADB_OD_Swe.SWE** beskrevet i avsnitt 3.1.3.1. De samme tekniske spesifikasjonene gjelder for dette subsettet.

Materialet finnes på to CD-er i arkivet; det er ikke funnet i andre lagringsformater. I likhet med resten av ADB_OD_Swe.SWE består det av en treningsdel og en testdel, som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Trening	swe4671	32	10	200	2 CD-er	0,6	-
Testing	swe4681	62	2	100			

Treningsmanuskriptet består av 20 vanlige setninger med to nølelyder i hver setning samt fire isolerte repetisjoner av hver nølelyd. Testmanuskriptet består av 50 vanlige setninger med to nølelyder i hver setning samt fire isolerte repetisjoner av hver nølelyd.

3.1.4 Dansk

3.1.4.1 Opptak for talegjenkjenning (ASR/Diktering), 16 kHz

Deldatabasen **ADB_OD_Dan.DKN** er samlet inn for produksjon av teknologi for, akustisk modellering for PC/Multimedia talegjenkjenning og for automatisk diktering (Office ASR and Dictation). Opptakene er gjort i lukket kontormiljø, og baserer seg på fonetisk balanserte manuskript, produsert på grunnlag av setninger fra NSTs danske korpus. Databasen består av en treningsdel og en testdel, der førstnevnte brukes til å trene selve den akustiske modellen, mens sistnevnte brukes for testformål. Fordelingen av opptak for de to delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Trening	dan0565	312	560	174720	181 CD-er + tape	61,3	Ursula
Testing	dan0611	987	56	55272	56 CD-er + tape	17,3	-

Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innspilt enhet, dvs. som oftest en setning eller noen tilfeller en frase eller enkeltord. Databasen er organisert etter samme katalogstruktur som beskrevet for norsk i avsnitt 3.1.2.1. Deldatabasen finnes ytterligere dokumentert på \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_OD.

Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	16 kHz
Oppløsning	16 bit
Format	Intel PCM
Kanaler	2 (stereo)

Dette formatet er i overensstemmelse med kravspesifikasjonene fra L&H. Materialets treningsdel finnes på serveren Ursula; \\Ursula\felles_ursu\opptak\adb_0565\speech. Denne deldatabasen er imidlertid bare kopiert og ikke konvertert til formatet som kreves for produksjon av IBM-baserte akustiske modeller. Dermed gjelder den samme tekniske spesifikasjon for dette lagringsformatet i dette tilfellet.

Innspillingsskriptet som treningsdataene bygger på har en dikteringsdel og en ASR-del. Førstnevnte del består av ordinære korpusekserperte setninger som kreves til generelle dikteringsformål, og utgjør manuskriptets 222 første enheter (setninger). Sistnevnte del omfatter de siste 90 enhetene og består av fraser består av fraser som utgjør personnavn, stedsnavn, enkeltord, akronymer og annet som kreves til spesifikke talegjenkjenningsformål (ASR).

Innspillingsskriptet som testdataene bygger på har en tilsvarende inndeling i en dikteringsdel og en ASR-del. Dikteringsdelen utgjør manuskriptets første 750 enheter, mens ASR-delen utgjør de siste 237 enhetene.

Hele materialet har blitt validert etter de kriterier og metoder som er nevnt i avsnitt 3.1.5. Valideringen har foregått i samarbeid med Center for Sprogteknologi ved Københavns Universitet. Valideringsfilene ligger på \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Filer\Dansk\ADB_OD_Dan.DKN\Validering\Diktering og \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Filer\Dansk\ADB_OD_Dan.DKN\Validering\ASR

3.1.4.2 Opptak for diktering, 22 kHz

Deldatabasen **ADB_D_IBM-D** er samlet inn for produksjon av teknologi for akustisk modellering for automatisk diktering (desktop). Opptakene spilt inn ved hjelp av IBM-programvaren ObjectRexx. Opptakene ble gjort i forbindelse med oppstart av samarbeidet mellom NST og IBM som ledd i opplæringsperioden av NST-ansatte.

Databasen består av tre deler innspilt til ulike formål, en testdel, en treningsdel og en modelleringsdel. Fordelingen av opptak i delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str (GB)	På server
Modellering	mod	260	109	28340	96 CD-er	6,5	Ursula
Testing	test	160	21	3360	inkl. kopi	0,6	-
Enrollment	enroll	155	21	3255		0,6	-

Opptakene fins i sin helhet i CD-arkivet. Det er ikke tatt sikkerhetskopi på streamertape av disse. Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	22 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	1 (mono)

Skriptene som opptakene er basert på ligger også på CD og har filnavn **da_mod_001.txt** (002, 003 etc), **da_enroll.txt** og **da_test.txt**. Disse finnes også på \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_D_IBM\Recordings\Innspillingsskript\Dansk. Opptakene er gjort i lukket kontormiljø, og baserer seg på fonetisk balanserte manuskript, produsert på grunnlag av nyhetstekst fra den NSTs danske korpus, dvs. avisen *Politiken*. Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innpilt enhet (setning, frase, enkeltord, tallrekke, bokstavrekke).

Som tabellen ovenfor viser, så finnes modelleringsdelen av denne deldatabasen på serveren Ursula, nærmere bestemt i katalogene

\\Ursula\felles_ursu\ac1\data\ddf (kvinner)
\\Ursula\felles_ursu\ac1\data\ddm (menn)

Den samme tekniske spesifikasjonen gjelder for denne kopien av datasettet. Her er lydfilene og manuskriptfilene lagret etter samme struktur som for den norske del av databasen:

\\Ursula\felles_ursu\ac1\data\ddf\a01\da\22.pcm (lydfiler)
\\Ursula\felles_ursu\ac1\data\ddf\a01\da\scripts\Word (manuskriptfiler)

Denne deldatabasen er ikke validert. Det foreligger derfor begrenset dokumentasjon, men en del informasjon finnes på \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_D_IBM

3.1.4.3 Opptak for telefoni, 8 kHz

Deldatabasen **ADB_T_Dan.DAN** inneholder opptak til mobiltelefoni; fastlinjeopptak er ikke representert. Disse er egnet til bruk ved produksjon av talegjenkjenningsteknologi for mobiltelefoni. NST har fulgt de generelle SpeechDat II-prosedyrene under opptak. Opptakene er gjort med L&Hs programvare ISDR.

Informantene har fått oppgitt telefonnummer og ringt inn og lest opp setninger. Opptakene inneholder 16 ytringer som er spontan tale i form av svar på spørsmål, og 40 ytringer med oppleste manuskriptsetninger.

Tabellen viser fordelingen av disse opptakene.

Formål	Skript	Linjer	Personer	Opptak	Lagring	Str GB	Server
Mobiltelefoni	scr0532A	56	1000	56000	10 CD-er + tape	3,5	Ursula
Mobiltelefoni	scr0532B	56					
Mobiltelefoni	scr0533	56					
Mobiltelefoni	scr8999	56					

Deldatabasen finnes i sin helhet i NSTs CD-arkiv. Det er foretatt fullstendig sikkerhetskopi på streamertape som ligger i safen. Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding	mu-Law
Filformat	wav
Samplingsfrekvens	8 kHz
Oppløsning	16 bit
Format	8 bit A-law compressed
Kanaler	1 (mono)

Databasen finnes også konvertert til IBM-format på serveren Ursula:

\\Ursula\felles_ursu\ambe\data\mdf (kvinner)
\\Ursula\felles_ursu\ambe\data\mdm (menn)

Her ligger dataene ordnet slik:

\\Ursula\felles_ursu\ambe\data\mdf\a01\da\8.pcm (lydfiler)
\\Ursula\felles_ursu\ambe\data\mdf\a01\da\scripts\Word (manuskriptfiler)

Følgende teknisk informasjon gjelder for denne deldatabasen i IBM-versjon:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	8 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	1 (mono)

Databasen er i sin helhet validert i henhold til kriteriene og metodene beskrevet i avsnitt 3.1.5. Valideringen har foregått i samarbeid med Center for Sprogteknologi ved Københavns Universitet. Foruten CD-arkivet ligger valideringsfilene også på \\Main\Felles\TIT\Ressourcer\ADB_oversigt\Filer\Dansk\ADB_T_Dan.DKN.

Deldatabasen er til en viss grad dokumentert i katalogen \\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_T, men her finnes kun begrenset med informasjon (kun en generert loggfil).

3.1.5 Validering

Begrepet ”validering” står sentralt i NSTs arbeid med akustiske ressurser. Felles valideringsprosedyrer er benyttet for alle innsamlingsprosjektene nevnt i avsnitt 3.1.2-3.1.4, og disse er i samsvar med prosedyrer for validering vedtatt av L&H.

Valideringen er foretatt av språkassistenter som har arbeidet tett sammen i grupper, under faglig koordinering av gruppeledere.

Valideringen innebærer gjennomlytting av hvert enkelt opptak, merking av taledelens utstrekning, kontroll av samsvar mellom ortografi og uttale, merking av ikke-verbale hendelser og bakgrunnsstøy, merking av feiluttale og dialektformer, og angivelse av opptakets generelle tekniske og lingvistiske kvalitet.

Ikke-verbale hendelser (”events”) angis ved et finitt sett av koder, definert som følger:

SPK	Angir tydelige lyder som ikke er tale, laget av taleren, som hosting, kremting, pusting, spyttlyder, osv. Markeres ikke dersom de forekommer ord-medialt.
FRA	Brukes ved feil uttale, repetisjoner eller ved forekomst av nonsensord. Markøren [FRA] plasseres foran ordet som er feil uttalt.
INT	Brukes ved tidsavgrensete lyder, som musikk, andre stemmer, sirener, osv. Markøren [INT] plasseres hvor lyden forekommer, ev. foran ordet ved ord-medial forekomst.
STA	Brukes ved gjennomgående lyd, f. eks. sirener. Markøren [STA] plasseres der lyden høres for første gang.
TRC	Brukes ved avkortete (trunkerte) setninger, enten i begynnelsen eller slutten av signal som er blitt av kortet. Dette kan forekomme hvis informant begynner å lese for tidlig, eller hvis en innringer legger på for tidlig.
FIL	Brukes ved fylte pauser, altså nølelyder. Markøren [FIL] plasseres hvor nølelyden forekommer i setningen.
DST	Brukes ved telefonforstyrrelser. Markøren [DST] plasseres foran ordet som blir forvrengt.
DIT	Brukes bare i tilfeller hvor en pipetone høres på begynnelsen av signalet. Denne er et signal til informanten om at han/hun kan begynne og snakke og skal ikke være en del av opptaket.

På grunnlag av disse annoteringene kan man ved modellering basert på databasene sørge for at dataene som inngår i den akustiske modellen kun inneholder menneskelig tale og ikke forstyrrende signaler, og dette er med på å høyne den akustiske modellens kvalitet.

Kvalitetsmerkingen av dataene baserer seg på observasjoner av både tekniske og lingvistiske forhold. En skala fra A til E er benyttet, der A er angir beste kvalitet og E angir forkastete opptak. Kvalitetskriteriene er for øvrig som følger:

A. Dersom man ikke hører andre ting enn talen gis karakteren A.
B. Dersom man observerer ikke-verbale hendelser i bakgrunnen (andre som snakker, spyttlyder, bakgrunnsstøy, osv.) gis karakteren B.
C. Dersom disse bakgrunnslydene er veldig tydelige, men ikke overdøver talen, gis karakteren C. C gis også dersom det er mindre enn 100 ms mellom begynnelsen eller slutten av talesignalet og begynnelsen eller slutten av opptaket.
D. Dersom bakgrunnslyder overdøver talen slik at det er vanskelig å høre hva informanten sier brukes karakteren D. D brukes også når feil uttale eller nølelyder forekommer. Dersom det er 0 ms mellom begynnelsen eller slutten av talesignalet og begynnelsen eller slutten av opptaket gis karakteren D.

E. Karakteren E fører til at opptaket blir forkastet. Dette vil være i tilfeller hvor man ikke forstår hva taleren sier, taleren leser feil ord, eller ingenting blir sagt. E brukes også hvis opptaket er avkortet.

Kontrollert innpust eller normal smatting regnes ikke som støy og er ikke markert som sådan. Det er heller ikke markert ved ubetydelig susing/knitring; litt lyd i bakgrunnen betraktes som normalt. Se avsnitt 3.1.7 for en mer detaljert beskrivelse av lingvistiske kvalitetskriterier.

For hvert sett av opptak er all tilhørende annotering samlet i en *Speech Logging File*. Dette er en ren tekstfil med filetternavn *.spl. Den inneholder fyllestgjørende metainformasjon om innspillingsutstyr, taleren, manuskriptet og de tilhørende enkeltfilene. En slik spl-fil inneholder informasjon om alle opptakene knyttet til et enkelt manuskript.

Eksempel på metainformasjon i spl-fil og informasjon om fire første opptak (av 320)[4]

```
[System]
Delimiter=>-<
Version=0001_1
CharacterSet=ANSI
ByteFormat=01
Script=463
Channels=2
Board=2;NI DSP2200
Frequency=16000
Coding=PCM;Linear
DOS Codepage=850
ANSI Codepage=1252
Memo=Kontor##1,5m, 2,5m, 1,5m, 2,5m##Shure bordmikrofon##Shur

[Info states]
1=Speaker ID>-<001>-<
2=Name>-<**** ****>-<
3=Age>-<57>-<
4=Sex>-<Female>-<
5=Region of Birth>-<Voss og omland>-<
6=Region of Youth>-<Voss og omland>-<
7=Remarks>-< frå hardanger>-<

[Session]
Directory=c:\adb_0463\data\scr0463\10\04631001\r4630001
Imported sheet file=c:\adb\dsdr\scripts\nor463\nor463.psh
Record session=1
Sheet number=1
RecDate=02 jul 1999
RecTime=08:52:13
Record duration=75' 36"
Number of recordings=312

[Record states]
1=2>-<>-<(...Vær stille under dette opptaket...)>-<1024>-<257024>-
<u0001001.wav>-<>-<1024>-<257024>-<bISa1>-<bISa1
2=2>-<>-<Tester en to tre fire fem seks sju åtte>-<1024>-<561024>-
<u0001002.wav>-<>-<257024>-<817024>-<tISa1>-<tISa1
3=2>-<>-<Blåbærturen ute på landet var en rein fornøyelse og flere av
```

[4] **** Angir anonymisering foretatt i denne rapporten.

```
turgåerene hadde kilosvis med bær.>-<1024>-<593024>-<u0001003.wav>-
<>-<817024>-<1409024>-<cISa1>-<cISa1
4=2>-<>-<Piloten hadde sitt svare strev med å få landet flyet i uvær
og svart natt.>-<1024>-<529024>-<u0001004.wav>-<>-<1409024>-
<1937024>-<cISa2>-<cISa2
5=2>-<>-<Det kan virke helt overveldende ute ved havet når den salte
skumsprøyten slår innover holmer og skjær.>-<1024>-<577024>-
<u0001005.wav>-<>-<1937024>-<2513024>-<cISa3>-<cISa3
```

Mer informasjon om validering finnes på
\\Main\felles\TIT\Ressourcer\ADB_oversigt\Dokumentasjon\ADB_T\Norsk\Validering

3.1.6 Dialektområder

I det følgende gis en oversikt over inndelingen i dialektområder som er brukt ved oppbygging av NSTs akustiske database. Denne ligger til grunn for datainnsamling av alle deldatabasene beskrevet ovenfor. For alle tre språk gjelder at talerne fordeler seg på aldersgruppene 18-70, og begge kjønn er representert. Det er ikke funnet noen samlet statistikk som angir fordelingen innenfor disse gruppene, men sporadiske funn i arkiverte rapporter antyder at fordelingen er noenlunde jevn. Dette bekreftes også av Kolbjørn Slethei ved Seksjon for Humanistisk Informatikk ved UiB, som tok del i utarbeidelsen av dialektinndelingen. For øvrig vil informasjon om informantfordeling uansett kunne la seg ekserpere fra spl-filene.

3.1.6.1 Norsk

Den norske databasen er fordelt på talere fra 11 dialektområder: Hedmark og Oppland, Oslo-området, Ytre Oslofjord, Sørlandet, Sør-Vestlandet, Bergen og Ytre Vestland, Voss og omland, Sunnmøre, Trøndelag, Nordland og Troms. Det foreligger ikke dokumentasjon som angir kriteriene for denne inndelingen, men Kolbjørn Slethei opplyser at denne er hovedsakelig basert på språklige vurderinger og sekundært på statistiske og sosioøkonomiske forhold. Et direktiv fra L&H tilsa at det maksimale antall dialektområder kunne være fem, så NST måtte "forhandle" seg frem til et høyere antall dialekter og dermed høyere antall informanter enn ved tilsvarende innsamlinger for andre europeiske språk.

3.1.6.2 Svensk

Den svenske databasen er fordelt på talere fra 10 dialektområder: Stockholm-området, Østre Sør-Sverige, Vestre Sør-Sverige, Västergötland, Vest-Sverige, Östergötland, Dalarna m/omegn, Göteborg-området, Midt-Sverige og Norrland. Denne inndelingen skiller seg fra den mer klassiske inndeling av svenske dialekter i sydsvenska mål, götamål, sveamål, gotländska mål, norrländska mål og östsvenska (f.eks. Mårtenson & Fjeldstad 1982). Ifølge en arkivert rapport er kriteriene for inndelingen "språklig homogenitet i kombination med befolkningstäthet och medelinkomst för kommunerna. … Inom varje region finner vi ett homogeniserat uttal." Det er altså lagt statistiske forhold til grunn, så vel som lingvistiske. Det later til at det er jevn geografisk fordeling mellom informantgruppene.

3.1.6.3 Dansk

Den danske databasen er fordelt på talere fra 7 dialektområder: København-området, Sjælland utenom København, Fyn, Nord-Jylland, Vest-Jylland, Øst-Jylland og Sør-

Jylland. Denne inndelingen skiller seg noe fra vanlig inndeling av danske dialekter i østdansk (Skåne, Bornholm), ødansk (Sjælland, Fyn, Sydøerne) og jysk, idet østdansk ikke er representert, mens sjællandsk er mer spesifikt inndelt. Det er ikke funnet dokumentasjon som angir hvorfor akkurat disse områdene ble valgt, men det er ikke grunn til å tro at kriteriene er annerledes enn de som ble brukt i den norske og svenske innsamlingen. Dermed er det antakelig statistiske og sosioøkonomiske kriterier som ligger til grunn for NSTs inndeling.

3.1.7 Språklige vurderinger (norsk)

I det følgende kommenteres en del språklige vurderinger og veivalg som er blitt gjort under arbeidet med den norske del av databasen.

For norsk (og svensk) gjelder at det opereres med et antall dialektområder og et antall informanter som relativt sett ligger høyere enn for tilsvarende innsamlingsprosjekt på øvrige språk i L&Hs portefølje. Dette har NST måttet rettferdiggjøre overfor sin teknologipartner med bakgrunn i den store dialektvariasjonen som forekommer i norsk.

Man har bevisst valgt å fokusere på bokmål og utelate nynorsk. Dette gir seg utslag i at alle innspillingsskript kun inneholder tekster på bokmål. Dersom taleren leser et bokmålsord som nynorsk, for eksempel hvis *skole* uttales som *skule* eller *åttende* uttales *åttande*, er ordets ortografi endret under valideringen slik at annotertingen stemmer overens med uttalen. I slike tilfeller er ordet merket som nynorsk i annotasjonsfilens kommentarfelt. Dette betraktes ikke som feil, og slike tilfeller er validert i henhold til kvalitetskriterium A, forutsatt at den ortografiske formen er innenfor nynorsk skriftnorm. En liknende strategi er valgt i tilfeller hvor en dialektform erstatter en bokmålsform, som når *venner* blir uttalt *venna*. Hvis dialektformen er relativt vanlig og ikke skiller seg særlig fra skriftnormen, vil kvalitetskriterium B være benyttet.

En forutsetning for arbeid med akustiske databaser er et tilhørende uttaleleksikon med informasjon om ordenes ortografi og uttale. NSTs uttaleleksikon er beskrevet i avsnitt 3.3. Dette leksikonet gjenspeiler uttaler som forekommer i den akustiske databasen. Det er tatt høyde for en viss grad av fonetisk variasjon i den akustiske databasen under utarbeidelsen av uttaleleksikonet. Leksikonet inneholder uttalevarianter i de tilfeller hvor de forekommer naturlig i opptakene og i talemål generelt; eksempelvis *vende* som [ˀvɛ.nə]/[ˀvɛn.də], og flere aksepterte uttaler av *morgen*, *måned*, *tredje*, *sytti* osv. Talerens faktiske uttale er ikke angitt i fonetisk representasjon i spl-filen men den er representert i leksikonet.

Fonetisk reduksjon er håndtert på ulikt vis, avhengig av konteksten den forekommer i. En del reduserte uttalevarianter vil få kvalitetsmerking A; dette gjelder varianter hvor reduksjon så å si alltid forekommer, som i *meteorolog* og *amerikaner* uttalt som [ˌmɛ.trʊ.ˈloːg] og [ˌɑm.ri.ˈkɑː.nər]. Dette regnes som den riktige uttalen av disse ordene, og i slike tilfeller vil faktisk den mer ortofone varianten betraktes som overartikulert og dermed få kvalitetsmerke B. Kvalitetsmerke B grunnet overartikulering gis også ved unaturlige geminater (*telefonnummer* uttalt med to *n*-er), eller ved bestemte nøytrumsformer hvor endelsen blir artikulert med plosiv (*hodet* uttalt [ˀhuː.dət]). Ved ikke-obligatoriske, men vanlige, reduksjoner vil

kvalitetskriterium B kunne være benyttet; eksempelvis *forutsette* uttalt [ˈfɔr.ʉ.ˌsɛ.tə] eller *Sarpsborg* uttalt uten [p]. Dette gjelder også f.eks. når *Hurdal* uttales [ˈhʉ.ˌɖɑɳ], hvor også annoteringens ortografi er endret til *Hurdalen*, i samsvar med uttalen.

Ved sjeldnere og uønskete reduksjoner brukes kvalitetsmerking fra C til E, avhengig av grad. Eksempelvis vil dialektformene *saukjan*, *akjan* og *nikkjan* få kvalitetsmerking E, og dermed bli forkastet som realisasjoner av tallordene.

Fonetisk/allofonisk variasjon forekommer naturlig i dataene og er annotert etter tilsvarende kvalitetskriterier. I de fleste tilfeller vil et sett av varianter være akseptert etter kvalitetskriterium A. Dette gjelder for eksempel variasjon av typen *kino* /çiː.nʊ/ vs. /ʃiː.nʊ/, *kort* /ˈkɔɯt/ vs. /ˈkɔʈ/ og *opp* /ˈɔp/ vs. /ˈʊp/. Normalt vil variasjonen være representert i form av fonemisk ulike varianter i uttaleleksikonet, men ikke i tilfellet *kino*, som kun er representert ved én uttaleform symbolisert ved [ç]. Konsekvensen av dette er at de to realisasjonene inngår som allofoniske varianter av fonemet [ç] i språkmodellene.

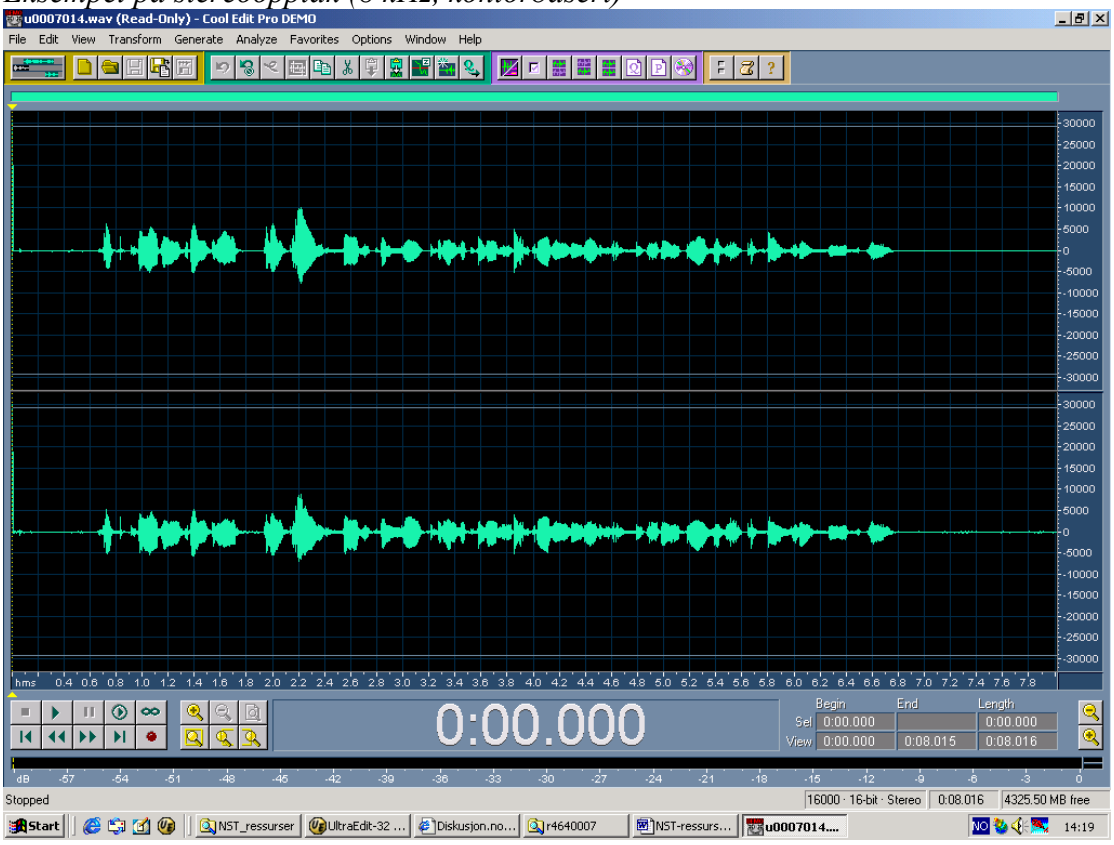
I mer dialektspesifikke tilfeller vil varianter kunne være merket med andre kriterier enn A. For eksempel er /b/d/g/ for [p/t/k] (bløte konsonanter) i sørlandsområdet annotert med kvalitetskriterium B, mens trøndersk /ʃ/ i verbformen *ser* ikke er godkjent, men kvalitetsmerket som E.

Det er grunn til å merke seg at databasen ikke er spesifikt konstruert for å representere fremmedspråklig uttale, men det kan forekomme informanter med utenlandsk opprinnelse i materialet. Instruksjonene er på dette feltet ikke entydige, men det later til at fremmedspråklig uttale godtas som norske uttalevarianter dersom den følger et konsekvent mønster, for eksempel bruk av /iː/ for [yː] (*lyd*) eller /uː/ for [ʉː] (*du*). Denne uttalen vil imidlertid ikke være merket med kvalitetskriterium A.

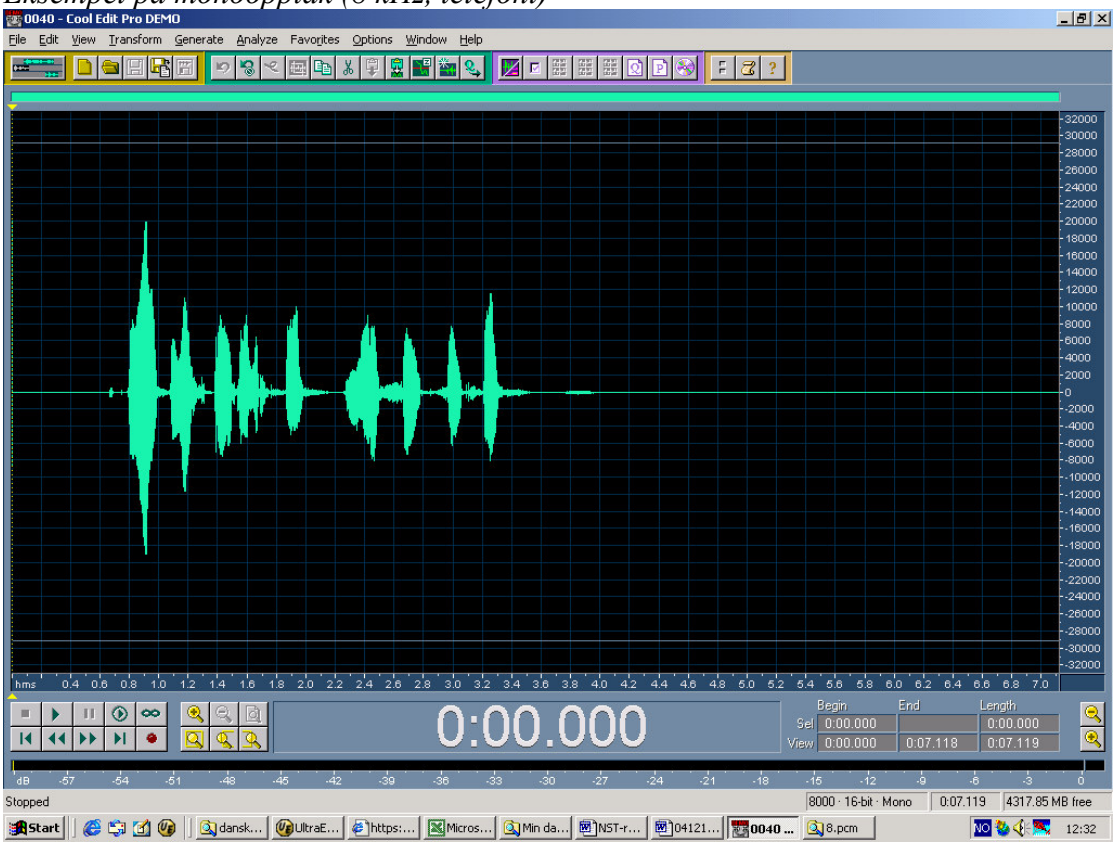
3.1.8 Kvalitetsvurdering

Stikkprøvekontrollen som er foretatt som del av gjeldende undersøkelse tyder på at lydmaterialet er av generelt høy kvalitet og grundig og nøyaktig validert. Lydkvaliteten er gjennomgående utmerket, med minimale mengder støy. Opptakene vil bli automatisk kuttet av innspillingsprogramvaren dersom talen overstiger et visst lydnivå. Ved stikkprøvekontrollen er det ikke funnet opptak hvor dette forekommer. Dette sees blant annet ved spektrogramvisning, hvor eventuelle kuttete opptak vil ha en flat kurve øverst, i stedet for å ha naturlige topper (jf. figurer på neste side).

Eksempel på stereoopptak (8 kHz, kontorbasert)



Eksempel på monoopptak (8 kHz, telefoni)



Som nevnt har man under innspillingen lagt inn 100-200 millisekunders tomrom mellom begynnelsen og slutten på talesignalet og begynnelsen og slutten på opptaket. Ved stikkprøvekontroll ble det funnet noen få telefoniopptak som var klippet på slutten, hvor siste del av setningen var utelatt. Problemets omfang er ikke klart, men det synes lite. Det er ikke funnet noen klippete opptak i opptakene fra kontormiljø. Uansett vil slike klippete opptak være fjernet ved hjelp av annotering under valideringen.

Kvalitetssikring av dataene gjennom validering later til å være grundig gjennomført. Valideringen fremstår som konsistent, noe som må tilskrives grundig dokumentasjon og konsekvent veiledning av språkassistenter. Informantenes tale er tydelig annotert ved hjelp av markører, den og lar seg skille fra uønsket lyd som innpust, utpust, spyttlyder, nøleøyder og annen støy. Dette gjør det mulig å produsere akustiske modeller som ikke inneholder støyforstyrrelser. Valideringen inneholder som nevnt også detaljerte opplysninger om eventuelle feiluttaler som talerne gjør.

Det må påtales at knirkestemme (*creaky voice*) forekommer i dataene uten at det er eksplisitt merket eller fjernet ved annotering. Dette er imidlertid ikke nødvendigvis noen svakhet, men kan inngå som en del av modelleringen da det også forekommer i naturlig tale. Under opptak for talesyntese er segmenter med knirkestemme fjernet eller overspilt.

Foruten fullstendig validering av dataene er det foretatt systematiske stikkprøve-kontroller, først av NSTs gruppeledere og deretter av L&Hs personale. Om lag 10 prosent av materialet er stikkprøvekontrollert av NST, mens 5 prosent er stikkprøvekontrollert i Belgia etter leveranse. Denne manuelle kontrollen overgår dermed de fem prosentene som ELRA-standarden krever. Apropos kvalitet og spotsjekking, i en intern møterapport fremgår følgende:

> NN fortalte først litt om situasjonen for dikteringsvalideringen. NST får ros for at vi har mye data, dvs. 900 opptak som er ferdigvalidert. I Belgia har de spotsjekket 150 dikteringsopptak og det er meget god kvalitet på disse. Derfor er det besluttet i Belgia at de dropper sin videre ”interne” spotsjekking, …

Ressursene fremstår som godt dokumenterte; jf. referanser med hyperlenker i de språkspesifikke avsnittene 3.1.2-3.1.4. Det er konsistens i navngiving av filer og katalogstruktur. Merkingen av CD-er ser ut til å være korrekt, selv om det ikke er foretatt en fullstendig opptelling av alt materialet som foreligger i CD-arkivet.

Det må fremheves at valideringsgruppene har hatt hyppige møter og har vært under faglig koordinering av lingvistisk skolerte gruppeledere. Valideringsarbeidet har vært samordnet mellom de tre språkene, og språkressursene har blitt bygget opp parallelt. Jevnlige møter innenfor og på tvers av de språkspesifikke gruppene gjør det klart at har NST gjort en betydelig innsats for å samordne arbeidet og utvikle en felles standard for valideringen. Dette har gitt materialet et entydig preg.

Det må imidlertid påpekes at en del av dokumentasjonen foreligger kun på norsk, noe som vil måtte endres dersom materialet skal inngå i en internasjonal ressursdatabase (jf. ELRAs kravspesifikasjon).

Oppbygging av akustiske databaser er ressurskrevende, og NST har lagt ned en stor innsats i oppgaven med å bygge opp en akustisk database bestående av i alt nærmere 2 millioner opptak. Bare i planleggingsfasen krever en slik ressursoppbygging omfattende oppgaver, som kartlegging av dialektområder, skriptkonstruksjon, prosjektstyring, ansettelser av personale til innsamlingen, opplæring, innkjøp av opptaksutstyr, transport av utstyr og personell til innspillingslokasjoner, rekruttering av informanter, kontrakter, logistikk osv. I neste fase kommer selve arbeidet med opptakene, før valideringen av foretas.

En naturlig innvending er at inndelingen i dialektområder ikke helt og holdent er gjort i henhold til dokumenterte dialektskiller, slik disse er beskrevet i litteraturen. Det virker som man har tatt praktiske og markedsmessige hensyn, mer enn hensyn til reelle dialektskiller i utvelgelsen av områder for innsamling av data. Man kan spørre seg hvorfor Voss er representert mens andre regioner med særpregete dialekter, som Sogn, Sunnfjord osv. ikke er tatt med. Det er imidlertid vanskelig å vurdere eventuelle konsekvenser dette valget har på reell gjenkjenningsrate, og om det er et mål å representere hver eneste dialekt i et så komplekst område som det nordiske. Viktigere er det å slå fast at et bredt geografisk område er representert i databasene for de tre språkene.

En annen betimelig innvending, som gjelder kun for norsk, er at nynorsk ikke er ivaretatt. Begge hovedmanuskriptene inneholder kun bokmålstekst. Det er presumptivt markedsmessige hensyn som ligger til grunn her, og språkpolitisk kan dette være problematisk. Riktignok representerer innsamlingen et stort geografisk område som inkluderer typiske nynorskområder, men det kan være fremtidig behov for å supplere materialet med nynorskbaserte opptak. For øvrig har NST valgt å håndtere forekomster av nynorskformer i den akustiske databasen på en forsvarlig måte.

Tross disse innvendingene kan det konkluderes med at NSTs akustiske database er i henhold til gjeldende standard slik denne er formulert etter ELRA-standarden, når det gjelder krav til teknisk kvalitet, språklig kvalitet, grad av validering, stikkprøve-kontroll, konsekvente metoder og dokumentasjon.

3.2 Akustiske databaser for talesyntese

NST har i flere omganger gjort opptak for produksjon av talesyntese, først i forbindelse med utvikling av skandinaviskspråklige versjoner av L&H-programvaren RealSpeak, som fremdeles er tilgjengelig fra selskapet Nuance, deretter for produksjon av IBMs talesyntese, som ikke rakk å nå markedet før NST gikk konkurs. Begge syntesene er konkatenative systemer; førstnevnte er en difonsyntese, mens IBM-syntesen er en datadrevet skjøtesyntese (unit-selection synthesis). I tillegg har NST utviklet IBM-baserte testsystemer for domenespesifikk syntese, under benevnelsen Phrase Splicing (frasespleising), samt gjort opptak for en del spesifikke kundeapplikasjoner.

Opptakene som ble gjort for utvikling av RealSpeak foregikk i sin helhet i L&Hs innspillingsstudio i Ieper, Belgia. Disse opptakene er ikke lokalisert i noen lagringsform på Voss. De er derfor ikke omtalt videre nedenfor.

3.2.1 IBMs talesyntese

I forbindelse med utviklingen av IBMs talesyntese ble det rekruttert profesjonelle stemmer, dvs. én herrestemme per språk. Opptakene er innspilt med IBM-programvare i et lydstudio på Voss, men proprietær innspillingsprogramvare er ikke til hinder for fremtidig bruk, da opptakene foreligger i anvendelig PCM-format. Følgende teknisk informasjon gjelder for de tre deldatabasene:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	44 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	2 (stereo): tale + laryngograf

Stereoopptakene har talesignal i én kanal, og signal fra laryngograf i andre kanal. Innspillingsmanuskriptene er basert på NSTs korpus. Et optimalisert utvalg av setninger er produsert ved hjelp av IBMs programvare OptScript. Manuskriptene er fonemisk optimaliserte for å oppnå bred dekning av mulige difonkombinasjoner. Manuskriptene er ikke prosodisk balanserte, men er likevel fordelt på kategorier som innebærer en viss prosodisk variasjon, som fortellende setninger, *hv*-spørsmål *ja*/*nei*-spørsmål og opplistinger. Et mindre subsett av manuskriptene inneholder setninger og tallformater som trengs i forbindelse med utvikling av en spesialisert taleapplikasjon for bankdomenet.

For de tre språkene fordeler ressursene seg som følger:

	Opptak og manuskriptfiler	**Antall opptak**	**Hv.av bank**
Norsk	\\Rene\felles_rene\TTS\TTS_NO\pcm\cs \\Rene\felles_rene\TTS\TTS_NO\pcm\cs\SCRIPTS	5363	417
Svensk	\\Rene\felles_rene\TTS\TTS_SW\sw_pcms\mf \\Rene\felles_rene\TTS\TTS_SW\sw_pcms\scripts	5279	267
Dansk	\\Heidi\felles_heidi\tts\TTS_DA\pcms\ca\all_rec \\Heidi\felles_heidi\tts\TTS_DA\pcms\ca\rec_scripts	4108	415

Det finnes sikkerhetskopi av dataene i CD-arkivet og på tape.

3.2.2 IBM Phrase Splicing

I tillegg til ovennevnte opptak beregnet på kommersiell utnyttelse ble det spilt inn flere datasett for produksjon av test- og demonstrasjonssystemer av IBMs programvare for frasespleising (*Phrase Splicing*), et system som er en hybrid mellom applikasjonsspesifikke innspillinger og vanlig konkatenativ talesyntese. Systemene er beregnet for bruk innenfor bankdomenet.

Følgende teknisk informasjon gjelder for disse deldatabasene:

Signalkoding	lineær PCM
Filformat	ukodete rådata (headerless raw)
Samplingsfrekvens	44 kHz
Oppløsning	16 bit
Format	Motorola PCM
Kanaler	2 (stereo): tale + laryngograf

Dette er de samme tekniske spesifikasjoner som for dataene beskrevet i avsnitt 3.2.1. Imidlertid er det en vesensforskjell når det gjelder materialets omfang og kvalitet. Disse opptakene er ikke gjort av profesjonelle stemmer, men informantene er i sin helhet rekruttert blant NSTs egne ansatte. Manuskriptene og antall opptak er også mindre enn for datasettene nevnt ovenfor. Det dreier seg om følgende ressurser fordelt på de tre språkene:

		Opptak og manuskriptfiler	Antall opptak	Bank
Norsk	kvinne	NB! Ikke lokalisert	2576	404
	mann	NB! Ikke lokalisert	2576	404
Svensk	kvinne 1	NB! Ikke lokalisert	1767	290
	kvinne 2	NB! Ikke lokalisert	1767	290
	mann	NB! Ikke lokalisert	1792	290
Dansk	kvinne	\\Heidi\felles_heidi\tts\TTS_DA\pcms\kni\	1917	415
	mann	\\Heidi\felles_heidi\tts\TTS_DA\pcms\lp\	1917	415

Merk at en del av disse dataene ikke er lokalisert på server eller i CD-arkiv. Dette gjelder kun selve PCM-filene; manuskript og annoteringsfiler ligger på \\Rene. Det er imidlertid god mulighet for at lydfilene finnes på streamertape, da det ligger sikkerhetskopi i arkivet av Linux-serverne hvor dataene ble overført etter opptak.

3.2.3 Opptak til spesifikke talestyrte applikasjoner (norsk)

Det er gjort en del mindre omfattende norske opptak for diverse talestyrte applikasjoner. Dette dreier seg bl.a. om opptak knyttet til det talestyrte sentralbordet AutoSwitch.

Disse dataene er ikke lokalisert på server eller i CD-arkiv og har derfor ikke vært gjenstand for undersøkelse. Dokumentasjon i form av informantavtaler viser at det for fem av informantene dreier seg om en helt ubetydelig datamengde, dvs. 3 opptak per informant, i tillegg til en usikker mengde opptak fra en kvinnelig norsk informant.

3.2.4 Kvalitetsvurdering

Opptakene gjort for utvikling av IBMs talesyntese, nevnt i 3.2.1, er gjort i lydstudio og med opptaktutstyr og laryngograf som er godkjent av IBM for produksjon av kommersielle talesyntesesystemer. Det er grunn til å understreke den høye samplingsfrekvensen, og den generelt høye kvaliteten på selve opptakene, samt at informantene er profesjonelle aktører. Opptakene har vært manuelt kontrollerte av NSTs produktutviklere for talesyntese. Disse akustiske databasene er segmentert og merket ved hjelp av IBMs annoteringsverktøy. Annoteringen er først gjort maskinelt og deretter manuelt kontrollert av utviklerne.

Materialet er i henhold til gjeldende standard for akustiske ressurser, og representerer ”state-of-the-art” for utvikling av talesyntese. Det er dermed anbefalt at inngår som del av en videreføring av NSTs akustiske database. Det vil også være svært gunstig dersom man fikk tilgjengeliggjort annoteringene i forbindelse med segmentering og parallellstilling (alignment) av lyd og tekst. Dette forutsetter antakelig en avtale med IBM på dette punktet.

Opptakene til frasespleising nevnt i avsnitt 3.2.2 er gjort som del av opplæring av NSTs ansatte i utvikling av syntseteknologi på IBMs avdelinger i Heidelberg, Hursley og Paris. Opptakene er gjort i til dels støyfulle kontormiljøer, og ikke i lydstudio. Dette gjør dem lite egnet til fremtidig bruk. For øvrig har disse databasene ikke samme fonetiske dekningsgrad som de som er omtalt i avsnitt 3.2.1.

Det er ikke dokumentert om opptakene til spesifikke talestyrte applikasjoner, nevnt i avsnitt 3.2.3, har den nødvendige tekniske og lingvistiske kvalitet som gjør dem egnet til fremtidig anvendelse. Imidlertid må det påpekes at disse er snevre hva angår bruksområde, fordi opptakene ble gjort for bestemte kunder (deriblant Voss kommune). De har dermed et setningstilfang som omfatter generelle setninger som *Hvem ønsker du å snakke med?*, men som også inkluderer et inventar av navn som er tilpasset kunden. Dermed vurderes ressursene nevnt i 3.2.3 som lite relevante for fremtidig bruk.

3.3 Leksikalske databaser

3.3.1 Generelt om NSTs leksikalske databaser

NSTs leksikalske databaser er utviklet og bearbeidet gjennom en periode fra bedriftens begynnelse og frem til konkursen. Utvikling av taleteknologisk programvare krever et leksikon bestående av ortografi og uttaleinformasjon og med et ordtilfang som dekker de manuskripter som de akustiske opptakene for talegjenkjenning og talesyntese og er basert på, samt et mer fullstendig vokabular av vanlige ord til bruk i taleteknologisk programvare.

Utgangspunktet for produksjon av NSTs leksikalske databaser har vært frekvensbaserte, ulemmatiserte ordlister som ble hentet ut fra NSTs norske, svenske og danske tekstkorpus. Første versjon av leksikonene var en såkalt 100k-liste, bestående av de 100.000 mest frekvente ordformer for hvert av språkene. Disse ble manuelt uttaletranskribert av NSTs egne transkriptører. For tilfellet dansk ble transkripsjonen foretatt ved Center for Sprogteknologi i København.

Språkteknikernes arbeid har i hovedsak bestått i manuelt å

- transkribere ordene i henhold til gjeldende konvensjoner for det aktuelle språk
- dekomponere ordene ved å sette inn + mellom sammensetningsledd og fugemorfem
- angi ordklasse for ordene
- angi ordklasse for sammensetningsleddene
- angi om ordet er et akronym eller forkortelse, og ev. ekspandere ordene
- duplisere leksikonposten hvis ordet er homograf.

Transkripsjonskonvensjonene er utarbeidet av lingvistisk skolerte gruppeledere. Selv om transkriptørene har arbeidet innenfor språkspesifikke grupper, er transkripsjonsarbeidet koordinert for de tre språkene, slik at metode og konvensjoner er standardisert på tvers av språkene.

Opprinnelig ble transkripsjonene gjort etter L&Hs proprietære transkripsjonssystem kalt L&H+. Etter bruddet med L&H ble transkripsjonene konvertert til det nøytrale fonetiske alfabetet SAMPA (Speech Assessment Methods Phonetic Alphabet; jf.

http://www.phon.ucl.ac.uk/home/sampa/index.html). Det er disse transkripsjonene som foreligger i den nåværende versjonen av leksikonene. Under IBM-samarbeidet ble et proprietært alfabet kalt CP5 (Common Phonology, version 5) benyttet, og et separat alfabet kalt Delta er utviklet spesifikt for TTS-utvikling. Med unntak av enkelte endringer i foneminventar, kommentert i de språkspesifikke avsnittene nedenfor, innebærer konverteringen kun en ren mapping fra ett tegnsystem til et annet. Konverteringen fra SAMPA til CP5/Delta innebærer ingen endringer i fonetisk inventar. Konverteringsverktøy mellom de ulike versjonene er ivaretatt. Katalogen \\Main\felles\TIT\Ressourcer\LDB\3.Trans.Conv_phon_tables inneholder transkripsjonskonvensjoner og tabeller over foneminventar som er brukt under ressursutviklingen.

I utgangspunktet er alle foreliggende transkripsjoner gjort manuelt. Ved utøking av materialet i ulike perioder er det imidlertid også gjort nytte av andre eksisterende ordboksressurser (blant annet norske NorKompLeks og det svenske Telia-materialet). I disse tilfellene har man konvertert ressursenes eksisterende transkripsjoner maskinelt til L&H+ eller SAMPA, avhengig av når utøkingen foregikk. For norsk og svensk er det utviklet en inflektor, dvs. et verktøy som genererer ortografier og transkripsjoner av bøyningsformer på grunnlag av et leksikon av grunnformer og deres transkripsjon. Disse genererte transkripsjonene har kun delvis vært gjenstand for manuell kontroll. Det fremgår i hver enkelt leksikonpost om formen er kun maskinelt generert eller også kontrollert av en transkriptør.

De tre leksikonene er navngitt etter følgende konvensjoner:
<språk><ååmmdd>NST.pron – eksempelvis: nor020110NST.pron.

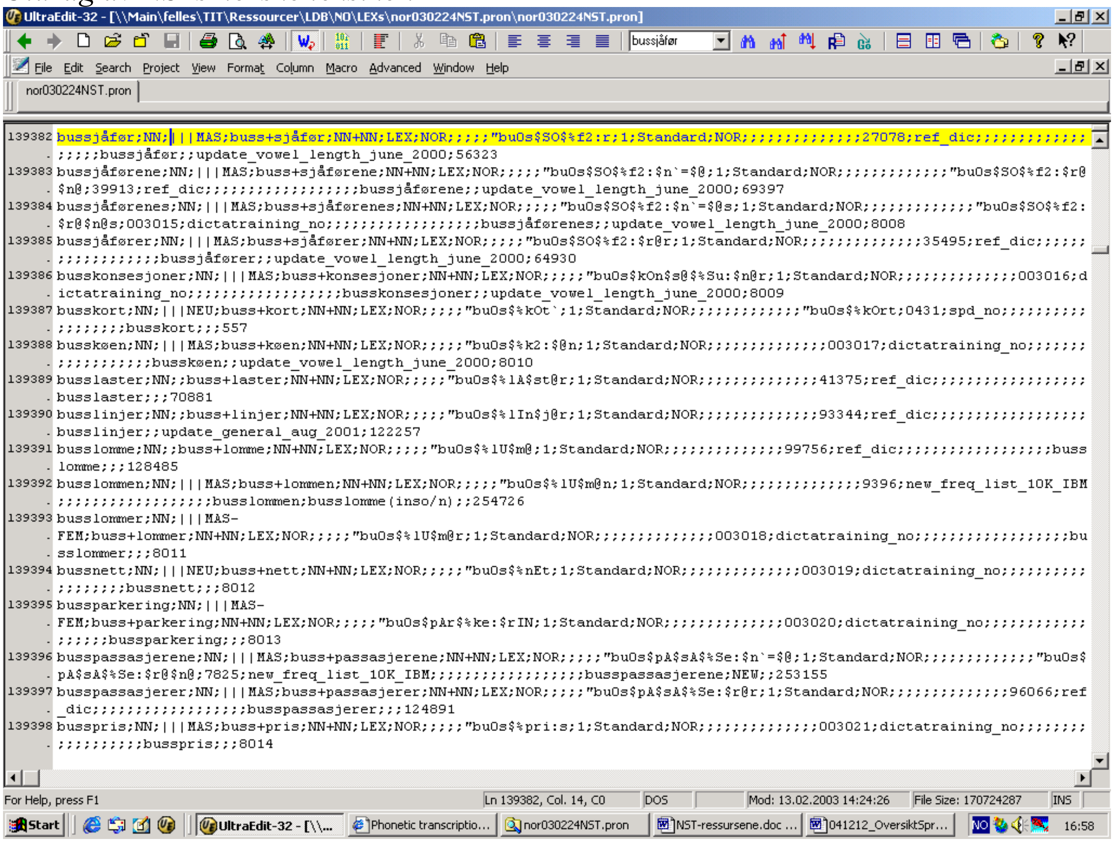
All informasjon om NSTs leksikalske databaser og selve databasene er samlet i en felles katalog: \\Main\felles\TIT\Ressourcer\LDB.

3.3.2 Databaseformat

NSTs leksikalske databaser foreligger som én tekstfil per språk. Denne består av én linje per leksikonpost, og informasjonen om leksikonposten foreligger i 51 felt, atskilt ved semikolon og nummerert i dokumentasjonen fra 0 til 50. Filen er alfabetisk sortert etter ortografi. Et eget dokument spesifiserer leksikonfilenes struktur og innhold: \\Main\felles\TIT\Ressourcer\LDB\4.NST_format_docs&tools\NST_lexicon_format.doc .

Et utsnitt av NSTs norske leksikon vises i figuren nedenfor.

Utdrag av NSTs norske leksikon



Informasjonen i leksikonet er ordnet hierarkisk, med inntil fem nivåer, hvorav kun de tre øverste er faktisk benyttet.

linjeskift	↵	skille mellom leksikonposter (ord); f.eks. mellom *landet* (v) og *landet* (n)
semikolon	;	skille på øverste nivå mellom ulike informasjonsfelt i leksikonposten; f.eks. mellom ortografi og transkripsjon
stolpe	|	skille på neste nivå; f.eks. mellom alternative transkripsjoner
bindestrek	-	skille på neste nivå (ikke benyttet)
komma	,	skille på laveste nivå (ikke benyttet)

De 51 informasjonsfeltene er beskrevet i detalj i dokumentasjonen, men i tabellen nedenfor gjengis den viktigste punktene herfra. Obligatoriske felter er merket i tabellen. I tilfelle et ord er merket som skrot (garbage) gjelder ikke regelen om obligatoriske felt.

Felt	Feltnavn	Obl.	Beskrivelse
0	orthography	X	Ordets ortografi, i korrigert form hvis nødvendig
1	extended pos	X	Ordklasseinformasjon med ev. underklassifisering av f.eks. proprier
2	morphology		Morfosyntaktisk informasjon (numerus, species, kasus, genus)
3	decomp	X	Dekomponering av sammensetninger
4	decpos	X	Ordklasseinformasjon for hvert dekomponeringsledd
5	source	X	Kilde; kodene LEX / INFL / COMP angir om leksikomposten stammer fra NSTs opprinnelige leksikon, inflektor eller annet verktøy (compounder)

6	language code orthography	X	Språkkode for ortografien, angir importord, fremmede navn osv; anvendte koder: NOR, NNO, DAN, SWE, ENG, FRE, GER, FIN, RUS, SPA, ITA, GRE, LAT, FOR
7	garbage		Angir om ordet er skrap
8	domain		Domene ordet er hentet fra; ikke anvendt men tenkt benyttet for felt som medisin, radiologi osv.
9	acronym/ abbreviation		Kodene ACR / ABBR angir om ordet er akronym eller forkortelse
10	expansion		Ekspansjon av felt 9 etter akronymets/forkortelsens betydning
11-26	transcriptions	X (11)	Fonetiske transkripsjoner, inkl. ev. ikke-forutsigbare varianter
27	automatically generated variants		Ev. genererbare uttalevarianter
28	set id	X	Unik numerisk identifikator for enkeltpost innenfor et transkripsjonsprosjekt
29	set name	X	Administrativ identifikator som angir navn på transkripsjonsprosjekt (f.eks. 100k)
30	style/status		Stilistisk informasjon
31	inflector role		Kodene BASE / INFLECTED angir om lekiskonposten har fungert som manuelt kontrollert grunnform eller er en maskinelt generert bøyningsform fra inflektoren
32	lemma		Lemmatilhørighet (ortografi og unik lemmakode) for ord fra inflektor
33	inflection rule		Regel som inflektoren har anvendt for å generere leksikonposten; brukes til å spore eventuelle feil i inflektoren
34	morph label		Morfologisk kode fra inflektor
35	compounder code		Prefikskode fra sammensettingsverktøy (ikke i bruk)
36	semantic info		Ordforklaring (fins i liten grad, kun i svensk fra Telia-listen)
37-45			Disponible felt
46	frequency		Frekvensopplysninger (ikke i bruk)
47	original orthography	X	Original ortografi, ulik felt 0 i korrigerte leksikonposter
48	comment field		Kommentarfelt
49	update info	X	Informasjon om leksikonpostens siste oppdatering/endring
50	unique id	X	Unik numerisk identifikator for leksikonposten

Leksikonets koder for syntaktisk og morfologisk informasjon er i henhold til Parole/SIMPLE-taggsettet; jf. http://www.ub.es/gilcub/SIMPLE/simple.html.

Morfosyntaktisk informasjon er lagt inn på to nivåer. Det finnes ordklasseinformasjon om alle leksikonposter, men detaljert morfosyntaktisk kode finnes kun for svenske og norske leksikonposter generert av NSTs inflektor. Denne er ordnet som følger i leksikonfeltene 1-2:

Felt 1	**Felt 2**
NN	numerus|species|kasus|genus
JJ	numerus|species|casus|genus|komparasjon
VB	aktiv/passiv|tempus/modus|annet
AB	komparasjon

Dekomponeringsfeltet inneholder markerte sammensetningsgrenser og fugemorfem angitt ved plusstegn, som i *uke+plan*. Følgende fugemorfemer anvendes på norsk:

s: mor+s+rollen
e: barn+e+hage
n: rose+n+kål
er: berlin+er+bolle
ar: laug+ar+dam
a: ferd+a+folk
me: lam+me+kotelett

Det er tatt høyde for en fremtidig mer detaljert morfologisk dekomponering. Skillet mellom stamme og suffiks er da tenkt angitt ved bruk av tilde, som følger: *bil+kjør~ing~en*.

Det er utarbeidet felles konvensjoner for når ord skal dekomponeres. Disse er semantisk og ikke etymologisk betinget. Sammensatte ord skal dekomponeres når hvert ledd har samme eller nærliggende betyding til den det får i sammensetningen. Videre er det en forutsetning for dekomponering at hvert enkelt ledd må kunne stå som enkeltord, dvs. at *tyttebær* og *hovedbruk* ikke deles, da verken *tytte-* eller *hoved-* kan fungere som enkeltord. Det er samsvar mellom dekomponering og uttale i den forstand at en morfologisk grense markert ved plusstegn alltid tilsvarer en stavingsgrense i transkripsjonen.

Flerordsuttrykk er merket som én ortografisk streng hvor en understrek tilsvarer mellomrommet i vanlig ortografi, som i *Ole_Irgens_vei*. Mellomrom forekommer ikke i leiksikonet. Prinsippene for leksikalsk dekomponering er for øvrig omtalt i hvert enkelt transkripsjonskonvensjonsdokument (jf. avsnitt 3.3.3-3.3.5).

I felt 30 angis eventuell informasjon om stilnivå og status. Dette forekommer særlig i den delen av det norske leksikonet som er generert av NSTs inflektor. Statusinformasjonen angir om leksikonposten er klammeform eller sideform i ordbøker. Følgende koder er anvendt for stilistisk koding:

Eksempel	Annotering
etter	Neutral
efter	Archaic|Klammeform
hoppa	Radical|Sideform
gutta	Casual|Klammeform
faen	Malediction

3.3.3 Norsk

NSTs norske leksikon ligger i tekstfilen \\Main\felles\TIT\Ressourcer\LDB\NO\LEXs\nor030224NST.pron\nor030224NST.pron. Katalogen på ett nivå opp, **..\LEXs** inneholder også tidligere versjoner av leksikonet. Følgende nøkkeltall gjelder for NSTs norske leksikon:

Antall poster i leksikonet	784 240	100,00 %
Antall skrotord (garbage)	2 282	0,29 %
Ord med minst én transkripsjon	753 621	96,10 %
Ord med to transkripsjoner	8 329	1,06 %
Ord med tre transkripsjoner	245	0,03 %

Ord med fire transkripsjoner	103	0,01 %
Sum av automatisk genererte transkripsjoner	244 264	
Totalt antall transkripsjoner	1 006 562	
Ord merket med ordklasseinformasjon	751 330	95,80 %
Ord merket med morfosyntaktisk kode	670 069	85,44 %
Ord merket med stilistisk informasjon	602 029	76,77 %
Manuelt kontrollert	254 419	32,44 %
Maskinelt generert av inflektor uten manuell kontroll	499 202	63,65 %

Differansen mellom totalt antall leksikonposter og ord med én fonetisk transkripsjon skyldes at enkelte såkalte garbage-poster (skrotord) kun merkes som sådan men får ingen øvrig annotering.

Arbeidet med en norsk inflektor er som nevnt inkorporert i det norske uttaleleksikonet. Inflektorens inndata består av 50 000 grunnformer; disse er i realiteten de samme som i leksikonet NorKompLeks, en innkjøpt ressurs som bygger på *Bokmålsordboka*. Grunnformene er konvertert til SAMPA og manuelt kontrollert og ev. endret i henhold til NSTs transkripsjonskonvensjoner. Om lag 254 000 ord er fra ulike prosjekter med manuell kontroll av transkripsjon, mens ca . 499 000 er stikkprøvekontrollerte, genererte leksikonposter fra inflektoren. Dette kan leses ut fra koden LEX eller INFL i leksikonets felt 5.

Samtlige ord, med unntak av skrotordene, er annotert med all informasjon merket som obligatorisk i tabellen ovenfor. Substantiver utgjør 67 prosent av leksikonet, adjektiver 14 prosent, verb 12 prosent, egennavn 7 prosent, adverb 0,1 prosent og øvrige grammatiske kategorier 0,1 prosent. Forkortelser og akronymer utgjør henholdsvis 310 og 940 leksikonposter.

Ordtilfanget er allment, og ingen spesialdomener er representerte. Leksikonet består av den frekvensbaserte 100k-listen og korresponderer med NSTs akustiske database på en slik måte at alle ordformer som fins i innspillingsmanuskriptene finnes transkribert i leksikonet. Videre inneholder leksikonet samtlige ord som finnes i *Bokmålsordboka* (via NorKompLeks/inflektor) med bøyningsformer, og samtlige ord som finnes i SpeechDat-materialet. Det er ved ulike delprosjekter lagt til egne subsett av personnavn, stedsnavn, bedriftsnavn, osv., blant annet hentet fra leksikonet Onomastica. Hvilket datasett ordformen hører til kan leses ut fra annoteringen i leksikonfelt 29. Kodene refererer til følgende datasett:

Datasett	Antall ord	Beskrivelse
ref_dic	101567	Frekvensbasert referanseordliste (100k)
namelex_no	34594	Navneleksikon som inneholder vanlige fornavn, doble fornavn, etternavn, stedsnavn (byer, steder, gater, land), hovedsakelig hentet fra Onomastica
dictatraining_no	31529	Ordtilfang fra treningsmanuskript for diktering
spd_no	4002	Ordtilfang fra SpeechDat
dictatest_no	3036	Ordtilfang fra testmanuskript for diktering
telephony_no	1985	Ordtilfang fra innspillingsskript for telefoni
asrtest_no	1715	Ordtilfang fra testmanuskript for ASR
asrtraining_no	1485	Ordtilfang fra treningsmanuskript for ASR
asrmobil_no	451	Ordtilfang innspillingsskript for telefoni
addon_feb2000_no	126	Diverse

addon_jan2000_no	31	Diverse
baseform_lex_no	44511	Grunnformer fra NorKompLeks
addon_acronyms_june_00	2	Diverse
addon_nsb_august_00	42	NSB stasjonsnavn og noen tallord
company_names_no	4304	Bedriftsnavn fra Onomastica
stock_names_no	108	Navn på børsnoterte bedrifter
no_13K	15816	Diverse

En tilhørende inspeksjonsfil **nor030224NST.pron_inspect.OUT** inneholder en mer detaljert kvantitativ fortegnelse over leksikonets innhold. Tekstfilen befinner seg i katalogen \\Main\felles\TIT\Ressourcer\LDB\NO\LEXs\nor030224NST.pron.

3.3.3.1 Transkripsjonskonvensjoner

NST har valgt Ålesund bymål som grunnlagsdialekt for sin leksikalske database. Dette ble gjort fordi den første talesyntesestemmen som bedriften utviklet var basert på denne dialekten. I den senere tid er det innført uttalevarianter i leksikonet som ikke er spesifikke for denne dialekten. Retningslinjer for fonetisk transkripsjon av NSTs norske leksikon er beskrevet i dokumentet \\Main\felles\TIT\Ressourcer\LDB\3.Trans.Conv_phon_tables\NO_trans_conv\TRANSCRIPTION _ALL_v1.5 rev.1.04.doc.

Foneminventaret som benyttes i leksikonet er dokumentert i \\Main\felles\TIT\Ressourcer\LDB\3.Trans.Conv_phon_tables\1.Phonetic_Tables\NO\PhonTable_norwegian_ipa_sampa_ibm_v.1.1.doc.

Transkripsjonen er i utgangspunktet fonemisk, og transkriptørene har blitt instruert i å "transkribere så breimaska som mulig". Prosedyrene for transkripsjon er for detaljerte til å bli gjennomgått her, men de viktigste punktene er følgende:

- stavelsesbærende konsonanter (nasal og lateral) er transkribert med egne symboler
- suprasegmentale angivelser som hovedtrykk, bitrykk og tonelag er markert
- vokallengder er eksplisitt markert
- bitrykk er markert i sammensatte ord
- flerordsuttrykk er markerte med fraseaksent i uttrykk som Ole_Irgens_vei[5] `/%u:$l@_¤"Ir$gEns_%v{*I/`
- halvlange vokaler er ikke representert ved egne symboler, men kan observeres som lange vokaler med bitrykk
- tjukk l er ikke anvendt
- retrofleks plosiv/nasal/lateral er representert ved egne symboler
- retroflekser er applisert syklisk, dvs. at et ord som *rapporten* transkriberes med en sekvens av retroflekser som følge av assimilasjon; jf. `/rA$"pO$t`n`=/`
- det angis ikke retrofleksering på tvers av sammensetningsgrense; jf. `/""vIn$t@r$%nAt/`
- det er innført symbolsk skille mellom /S/ som i *sju* og /s`/ som i *kors*, selv om de to har sammenfallende uttale og inngår i de samme akustiske modeller. Dette er gjort for å forenkle innføring av genererte uttalevarianter uten retrofleks for dialekter hvor dette ikke forekommer (som bergensk).

[5] SAMPA-transkripsjoner er gjengitt her.

Stavelsesgrenser er markert etter et moderat Maximal Onset-prinsipp (MOP). Dette innebærer at man har maksimert opptakter innenfor språkets fonotaks. Sammensetningsgrense blokkerer systematisk for MOP-prinsippet; altså et ord som *Drammensveien* har stavingsgrense foran *-veien*, og fuge-/s/ er ikke en del av etterleddets første stavelse; jf. /"drA.m@ns.'2v {*I.@n/.

I tillegg til en del manuelt innlagte alternative uttaler for uforutsigbare varianter som *tredje* (/""tre:$dj@/ vs. /""tr{*I$@;/), *sytti* (/"s9$tI/ vs. /"sY$tI/) osv., er det lagt inn maskinelt genererte uttalevarianter hvor disse lar seg systematisk forutsi. Dette omfatter blant annet variasjon knyttet til retrofleksering og bruk av stavelsesbærende konsonant, og variasjon i vokalkvalitet i ord som *oppgave* (/""Op$%gA:$v@/ vs. /""Up$%gA:$v@/). En konsekvens av dette er at et ord som *rapporten* får et sett av genererte uttalevarianter:

/rA$"pOr$t@n/
/rA$"pU$t`n`=/
/rA$"pUr$t@n/
/rA$"pOr$t@n/
/rA$"pU$t`n`=/
/rA$"pUr$t@n/

Konverteringen fra det opprinnelige L&H-formatet til SAMPA har innebåret noen endringer i fonetisk inventar, særlig knyttet til realisasjon av fonemet [e]. Mens alle realisasjoner var samlet i ett fonem i det L&H-baserte systemet ble det innført egne symboler for /e:/, /ɛ/ og /ə/ illustrert ved *lekselesing* /""lEk$s@$%le:$sIN/ i SAMPA-versjonen av leksikonet.

3.3.3.2 Innkjøpte leksikalske ressurser

I tillegg til det egenutviklete leksikonet har NST anskaffet eksterne norske leksikalske ressurser, dels ved kjøp, dels som del av tingsinnskudd gjennom aksjeemisjon (såkalte "in-kind"-anskaffelser). Dette omfatter følgende leksikalske ressurser:

Navn/Plassering	Innhold	Annotering
INSO \\Main\felles\TIT\Ressourcer\OLD_LDB\felles\INSO\INSO-inflector\languages\bokmål-utpakket \\Main\felles\TIT\Ressourcer\OLD_LDB\felles\INSO\INSO-inflector\languages\Nynorsk-utpakket	71006 grunnformer 595619 bøyningsformer	bøyningskode, ordklasse, morfologisk kode, sammensetnings-informasjon
Norkompleks \\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\NORKOMPLEKS	80443 grunnformer 460777 bøyningsformer	ordklasse, morfologisk kode, fonetisk transkripsjon
Onomastica \\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\ONOMASTICA	556499 navn, hvorav etternavn: 82416; fornavn: 239393; doble fornavn: 27096; stedsnavn: 6180; gatenavn: 88161; bedriftsnavn: 100179; utlandske navn: 13074	ordklasse, fonetisk transkripsjon, kvalitetsmerking

Statistisk sentralbyrå \\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\NAVN	71795 navn, hvorav etternavn: 56808; fornavn: 6639 (m) + 7097 (k); stedsnavn: 1251	frekvens, ordklasse
Bronnoy_navn \\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\Bronnoey_navn	1019643 navn hvorav bedriftsnavn: 791585; fornavn: 139856; mellomnavn: 23672; etternavn: 64530	ordklasse

Som nevnt ovenfor er deler av disse ressursene inkorporert i NSTs norske leksikon, og det er dermed sammenfallende ordtilfang mellom disse ressursene.

3.3.3.3 Leksikalske verktøy

Det er bygget opp en god del språkbehandlingsverktøy som automatiserer og forenkler arbeidet med leksikalske ressurser. Disse omfatter følgende:

Verktøy for leksikoninspeksjon
Dette verktøyet sjekker at leksikonet inneholder all obligatorisk informasjon, kontrollerer at transkripsjonene kun inneholder valide tegn, samt lager statistikk.
\\Main\felles\TIT\Ressourcer\LDB\NO\LEXs\nor030224NST.pron\inspect_lex.pl

Inflektor
Inflektoren genererer bøyningsformenes ortografi og transkripsjoner på grunnlag av et grunnformsleksikon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\INFLEKTOR

Ordsammensettingsverktøy (Recompounder)
Verktøyet prosesserer sammensetninger og tildeler bitrykk, trykkforskyvinger osv. på grunnlag av inndata bestående av enkeltleddenes ortografi og transkripsjon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\RECOMPOUNDER_NO

Tallordgenerator
Verktøyet konverterer norske tallord fra siffernotasjon til ortografi.
\\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\TALLORD

Transkripsjonskorreksjonsprogram
Verktøyet gjør en grunnleggende kontroll av fonetiske transkripsjoner i leksikonet.
\\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\TOOLS\check

Fonotaktisk parser
Verktøyet gjør en mer omfattende kontroll av fonetiske transkripsjoner i leksikonet.
\\Main\felles\TIT\Ressourcer\OLD_LDB\norsk\RESURSER_NO\TOOLS\Sofias_parser

Verktøy for grafem-til-fonem-konvertering (G2P)
Verktøyet konverterer ord fra ortografi til fonetisk representasjon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\felles\IBM_PREPARE\SITE_FELLES\FELLES_CD1\Felles_g2pLight\No.zip

Dekomponeringsverktøy
Verktøyet deler opp sammensetninger på grunnlag av lister over sammensetningsledd.
\\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\DECOMPOUNDERS\nor

Dekomponeringsverktøy (pilotversjon)
Verktøyet deler opp sammensetninger på grunnlag av konsonantklustermetode og lister over vanlige sammensetningsledd.
\\Server2\ABLEX\ABLEKS\tool

3.3.4 Svensk

NSTs svenske leksikon ligger i tekstfilen
\\Main\felles\TIT\Ressourcer\LDB\SW\LEXs\swe030224NST.pron\swe030224NST.pron
Katalogen på ett nivå opp, **..\LEXs** inneholder også tidligere versjoner av leksikonet.
Følgende nøkkeltall gjelder for NSTs svenske leksikon:

Antall poster i leksikonet	927 167	100,00 %
Antall skrotord (garbage)	1 979	0,21 %
Ord med minst én transkripsjon	927 167	100,00 %
Ord med to transkripsjoner	6 213	0,67 %
Ord med tre transkripsjoner	204	0,02 %
Ord med fire transkripsjoner	51	0,01 %
Sum av automatisk genererte transkripsjoner	0	
Totalt antall transkripsjoner	933 635	
Ord merket med ordklasseinformasjon	925 821	99,85 %
Ord merket med morfosyntaktisk kode	852 718	91,97 %
Ord merket med stilistisk informasjon	0	0,00 %
Manuelt kontrollert	249 901	26,95 %
Maskinelt generert av inflektor uten manuell kontroll	677 266	73,05 %

Arbeidet med en svensk inflektor er inkorporert i det svenske uttaleleksikonet. Inflektoren er basert på Telia-materialet, som er en innkjøpt ressurs. Grunnformene herfra er konvertert til SAMPA og manuelt kontrollert og ev. endret i henhold til NSTs transkripsjonskonvensjoner. Om lag 250 000 ord i leksikonet er fra ulike prosjekter med manuell kontroll av transkripsjon, mens ca . 677 000 er stikkprøvekontrollerte genererte leksikonposter fra inflektoren.

Samtlige ord i leksikonet, med unntak av skrotordene, er annotert med informasjon i alle obligatoriske felt. Substantiver utgjør 63 prosent av leksikonet, verb 16 prosent, adjektiver 14 prosent, egennavn 5 prosent, adverb 0,3 prosent og øvrige grammatiske kategorier 0,2 prosent. Forkortelser og akronymer utgjør henholdsvis 100 og 452 leksikonposter.

Ordtilfanget er allment, og ingen spesialdomener er representerte. Leksikonet består av den frekvensbaserte 100k-listen og korresponderer med NSTs akustiske database på en slik måte at alle ordformer som fins i innspillingsmanuskriptene finnes transkribert i leksikonet. Videre inneholder leksikonet samtlige ord som finnes i de innkjøpte Telia-materialet og samtlige ord som finnes i SpeechDat-materialet. Det er ved ulike delprosjekter lagt til egne subsett av personnavn, stedsnavn, bedriftsnavn,

osv. Hvilket datasett ordformen hører til kan leses ut fra annoteringen i leksikonfelt 29. Kodene refererer til følgende datasett:

Datasett	Antall ord	Beskrivelse
enter_se	86043	Frekvensbasert referanseordliste (100k)
teliabaseforms_se	73184	Grunnformer fra Telia-leksikonet
telia_personnames_se	13000	Personnavn fra Telia-leksikonet
spdnew_se	3764	Ordtilfang fra SpeechDat-materialet
dictatraining_se	34941	Ordtilfang fra treningsmanuskript for diktering
asrtraining_se	546	Ordtilfang fra treningsmanuskript for ASR
dictatest_se	2786	Ordtilfang fra testmanuskript for diktering
asrtest_se	768	Ordtilfang fra testmanuskript for ASR
lacking_se_17may	8	Diverse
spd_se	20201	Ordtilfang fra SpeechDat-materialet
telia_placenames_se	16560	Stedsnavn fra Telia-leksikonet
finnsws_se	2443	Ordtilfang fra SpeechDat-materialet
guru_dict	2506	Ordtilfang fra svensk dikteringsprosjekt

En tilhørende inspeksjonsfil **swe030224NST.pron_inspect.OUT** inneholder en mer detaljert kvantitativ fortegnelse over leksikonets innhold. Tekstfilen fins i katalogen \\Main\felles\TIT\Ressourcer\LDB\SW\LEXs\swe030224NST.pron

3.3.4.1 Transkripsjonskonvensjoner

Transkripsjonene er gjort etter ”centralsvenskt standardspråk” med utgangspunkt i Stockholm-dialekten. Transkripsjonene er basert på de samme prinsippene som beskrevet for norsk i avsnitt 3.3.3.1 med hensyn til retrofleksering, suprasegmentale angivelser, MOP-prinsippet osv. For øvrig vises til følgende retningslinjer for fonetisk transkripsjon av NSTs svenske leksikon:
\\Main\felles\TIT\Ressourcer\LDB\3.Trans.Conv_phon_tables\SW_trans_conv\temp_Swedish_transcriptionconventions_SAMPA_ver02.doc

Foneminventaret er beskrevet i følgende dokument
\\Main\felles\TIT\Ressourcer\LDB\3.Trans.Conv_phon_tables\1.Phonetic_Tables\SW\PhonTable_swedish_ipa_sampa_ibm_v.1.1.doc

3.3.4.2 Innkjøpte leksikalske ressurser

I tillegg til det egenutviklete leksikonet har NST anskaffet enkelte svenske leksikalske ressurser, dels ved kjøp, dels som del av tingsinnskudd. Dette omfatter det følgende leksikalske ressurser:

Navn/Plassering	Innhold	Annotering
INSO \\Main\felles\TIT\Ressourcer\OLD_LDB\felles\INSO\INSO-inflector\languages\Swedish-utpakket	53205 grunnformer 423581 bøyningsformer	bøyningskode, ordklasse, morfologisk kode, sammensetnings-informasjon
Telia-materialet \\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\RESURSER_SE\TELIAORIGINAL	116190 grunnformer 907000 bøyningsformer	ordklasse, morfologisk kode, fonetisk transkripsjon

Telia navneleksikon \\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\RESURSER_SE\TELIAORIGINAL	171926 navn, hvorav stedsnavn: 5967; gatenavn: 33107; etternavn: 122181; fornavn: 10671	ordklasse, fonetisk transkripsjon, frekvens
Enter-listen \\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\RESURSER_SE\LEX\PRONUNCIATION\ORG\enter.ref	122314 ordformer	fonetisk transkripsjon
Postvesenet (filen ord.gatunamn.posten.txt) IKKE FUNNET	176743 stedsnavn	ingen
Medisinske ordlister \\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\MEDICIN_se\MEDICIN	Uvisst antall	fonetisk transkripsjon 2700 ord innenfor radiologi, del av pilotprosjekt
Terminologi og nomenklatur-centralen, TNC \\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\RESURSER_SE\TNC	Termbase med 24109 termposter fra 21 ulike fagspråklige områder ordlister	Ordklasse, bøyningsinfo, definisjon, relaterte termer, språklige eksempler, domene, oversettelser

3.3.4.3 Leksikalske verktøy

Det er bygget opp ulike språkbehandlingsverktøy som automatiserer og forenkler arbeidet med leksikalske ressurser. Disse omfatter følgende:

Verktøy for leksikoninspeksjon
Dette verktøyet sjekker at leksikonet inneholder all obligatorisk informasjon, kontrollerer at transkripsjonene kun inneholder valide tegn, samt lager statistikk.
\\Main\felles\TIT\Ressourcer\LDB\SW\LEXs\swe030224NST.pron\inspect_lex.pl

Inflektor
Inflektoren genererer bøyningsformenes ortografi og transkripsjoner på grunnlag av et grunnformsleksikon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\SVENSK_INFLEKTOR

Dekomponeringsverktøy
Verktøyet deler opp sammensetninger på grunnlag av bokstavkombinasjoner og lister over hyppige sammensettingsledd.
\\Main\felles\TIT\Ressourcer\OLD_LDB\Svensk\DECOMPOUNDERS\swe

Ordsammensettingsverktøy (Recompounder)
Verktøyet prosesserer sammensetninger og tildeler bitrykk, trykkforskyvinger osv. på grunnlag av inndata bestående av enkeltleddenes ortografi og transkripsjon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\felles\IBM_PREPARE\SITE_FELLES\FELLES_CD1\Felles_recompounder\RECOMPOUNDER_SE_3002.zip

Transkripsjonskorreksjonsprogram
Verktøyet gjør en grunnleggende kontroll av fonetiske transkripsjoner i leksikonet.
\\Main\felles\TIT\Ressourcer\LDB\6.Felles_tools\CorrPhon

Verktøy for grafem-til-fonem-konvertering (G2P)
Verktøyet konverterer ord fra ortografisk til fonetisk representasjon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\felles\IBM_PREPARE\SITE_FELLES\FELLES_CD1\Felles_g2pLight\Se.zip

Tallordgenerator
Verktøyet konverterer svenske tallord fra siffernotasjon til ortografi.
\\Main\felles\TIT\Ressourcer\LDB\SW\Generelle scripts\TTS\Digits\siffror.pl

3.3.5 Dansk

NSTs danske leksikon ligger i tekstfilen
\\Main\felles\TIT\Ressourcer\LDB\DA\LEXs\dan030224NST.pron\dan030224NST.pron
Katalogen på ett nivå opp, **..\LEXs** inneholder også tidligere versjoner av leksikonet.
Følgende nøkkeltall gjelder for NSTs danske leksikon:

Antall poster i leksikonet	237 873	100,00 %
Antall skrotord (garbage)	450	0,19 %
Ord med minst én transkripsjon	237 873	100,00 %
Ord med to transkripsjoner	13 710	5,76 %
Ord med tre transkripsjoner	1 370	0,58 %
Ord med fire transkripsjoner	0	0,00 %
Sum av automatisk genererte transkripsjoner	0	
Totalt antall transkripsjoner	252 953	
Ord merket med ordklasseinformasjon	236 307	99,34 %
Ord merket med morfosyntaktisk kode	0	0,00 %
Ord merket med stilistisk informasjon	0	0,00 %
Manuelt kontrollert	237 873	100,00 %
Maskinelt generert av inflektor uten manuell kontroll	0	0,00 %

For dansk er det ikke utviklet en inflektor, og samtlige leksikonposter er dermed manuelt transkribert. Samtlige ord i leksikonet, med unntak av skrotordene, er annotert med informasjon i alle obligatoriske felt. Substantiver utgjør 52 prosent av leksikonet, egennavn 24 prosent, adjektiver 12 prosent, verb 10 prosent, adverb 1,9 prosent og øvrige grammatiske kategorier 0,4 prosent. Forkortelser og akronymer utgjør henholdsvis 740 og 247 leksikonposter.

Ordtilfanget er allment, og ingen spesialdomener er representerte. Leksikonet består av en frekvensbasert 100k-liste og korresponderer med NSTs akustiske database ved at alle ordformer som fins i innspillingsmanuskriptene finnes transkribert i leksikonet. Videre inneholder leksikonet samtlige ord som finnes i det danske INSO- og SpeechDat-materialet. Det er ved ulike delprosjekter lagt til egne subsett av personnavn, stedsnavn, bedriftsnavn, osv.. Hvilket datasett ordformen hører til kan leses ut fra annoteringen i leksikonfelt 29. Kodene refererer til følgende datasett:

Datasett	Antall ord	Beskrivelse
ref.dic	112 999	Frekvensbasert referanseordliste (100k)
spd_da	16 547	Ordtilfang fra SpeechDat-materialet
tel_da	22 978	Ordtilfang fra innspillingsskript for telefoni
off_da	4 106	Ordtilfang fra innspillingsmanuskript for diktering
nml_da	29 418	Navneleksikon
inso_da	40 125	Grunnformer fra INSO-materialet
stn_da	9 125	Gatenavn fra Krak-materialet
lstn_da	2 211	Etternavn
kons_da	109	Diverse

En tilhørende inspeksjonsfil **dan030224NST.pron_inspect.OUT** inneholder en mer detaljert kvantitativ fortegnelse over leksikonets innhold. Tekstfilen befinner seg i katalogen \\Main\felles\TIT\Ressourcer\LDB\DA\LEXs\dan030224NST.pron.

3.3.5.1 Transkripsjonskonvensjoner

Transkripsjonene tar utgangspunkt i Københavnsdialekten og er basert på de samme prinsippene som beskrevet for norsk i avsnitt 3.3.3.1 med hensyn til suprasegmental markering og MOP-prinsippet. For øvrig vises til følgende retningslinjer for fonetisk transkripsjon av NSTs danske leksikon:
\\Main\felles\TIT\Ressourcer\LDB\3.Trans.Conv_phon_tables\DA_trans_conv\DA_SAMPA_transkonv..doc

Merk at betydelig plass er viet plass er viet håndteringen av stød og /r/-diftongering og vokalrealisasjoner i /r/-kontekst.

Foneminventaret er beskrevet i følgende dokument:
\\Main\felles\TIT\Ressourcer\LDB\3.Trans.Conv_phon_tables\1.Phonetic_Tables\DA\PhonTable_danish_ipa_sampa_ibm_v.1.3.doc

3.3.5.2 Innkjøpte leksikalske ressurser

I tillegg til det egenutviklete leksikonet har NST anskaffet enkelte danske leksikalske ressurser. Dette omfatter følgende:

Navn/Plassering	Innhold	Annotering
INSO \\Main\felles\TIT\Ressourcer\OLD_LDB\felles\INSO\INSO-inflector\languages\danish-utpakket	75000 grunnformer 520000 bøyningsformer	bøyningskode, ordklasse, morfologisk kode, sammensetnings-informasjon
Institut for Navneforskning \\Main\felles\TIT\Ressourcer\OLD_LDB\dansk\OLD_LDB_DA\ProperNames\Navneleksikon.zip	240000 for-, etter-, steds- og personnavn.	
KRAK forlag \\Main\felles\TIT\Ressourcer\OLD_LDB\dansk\OLD_LDB_DA\ProperNames\Krak.zip	100 000 gatenavn	

3.3.5.3 Leksikalske verktøy

Det er bygget opp færre leksikalske språkbehandlingsverktøy for dansk enn for de øvrige språkene. Følgende verktøy foreligger i NSTs database.

Verktøy for leksikoninspeksjon
Dette verktøyet sjekker at leksikonet inneholder all obligatorisk informasjon, kontrollerer at transkripsjonene kun inneholder valide tegn, samt lager statistikk.
\\Main\felles\TIT\Ressourcer\LDB\SW\LEXs\swe030224NST.pron\inspect_lex.pl

Ordsammensettingsverktøy (Recompounder)
Verktøyet prosesserer gitt sammensetninger og tildeling av bitrykk, trykkforskyvinger osv. på grunnlag av inndata bestående av enkeltleddenes ortografi og transkripsjon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\dansk\OLD_LDB_DA\COMPOUNDER\DK

Verktøy for grafem-til-fonem-konvertering (G2P)
Verktøyet konverterer ord fra ortografi til fonetisk representasjon.
\\Main\felles\TIT\Ressourcer\OLD_LDB\felles\IBM_PREPARE\SITE_FELLES\FELLES_CD1\Felles_g2pLight\Dk.zip

Transkripsjonskorreksjonsprogram
Verktøyet gjør en grunnleggende kontroll av fonetiske transkripsjoner i leksikonet.
\\Main\felles\TIT\Ressourcer\LDB\6.Felles_tools\CorrPhon

3.3.6 Kvalitetsvurdering

Leksikalske ressurser som dem beskrevet i foregående avsnitt er nødvendig for taleteknologiske applikasjoner. Tilgjengelighet til NSTs norske, svenske og danske leksikon er langt på vei en forutsetning for å kunne nyttiggjøre seg den akustiske databasen.

NSTs egenutviklete leksikon fremstår som omfattende og godt dokumenterte (jf. hyperlenker til dokumentasjon i foregående avsnitt). Det er lagt ned en betydelig innsats i å transkribere ord manuelt og annotere dem med informasjon om ordklasse, sammensetninger osv. Dette leksikalske arbeidet har vært ledet av gruppeledere som har samordnet innsatsen på tvers av de tre språkene.

Alle tre leksikonene består av et grunnlagsmateriale på ca. 250 000 manuelt transkriberte ordformer. Det norske og svenske leksikonet er supplert med genererte ordformer utarbeidet av en inflektor (henholdsvis 500 000 og 677 000 ordformer). Det må understrekes at dette materialet ikke har vært gjenstand for manuell kontroll, mens det danske leksikonet i sin helhet har det.

Den leksikalske databasen er tilpasset den akustiske databasen ved at dens innhold sammenfaller med ordtilfanget i innspillingsmanuskriptene. Ordtilfanget i leksikonet er imidlertid langt mer omfattende enn som så. Leksikonene har vært utvidet ved flere anledninger og har en bred dekning av allment vokabular. Det omfatter forholdsvis mange navn, inkludert fornavn, doble fornavn, etternavn, stedsnavn, gatenavn, byer, land, stasjonsnavn, bedriftsnavn, osv.

Imidlertid er det en svakhet at dette arbeidet ikke har vært videreført de siste årene. Dermed dekker ikke ordtilfanget nyere neologismer og importord. For eksempel er et ord som *tsunami* er ikke representert i det norske leksikonet. Ingen særskilte fagområder er representert, og ved applikasjon innenfor spesifikke fagfelt, som for eksempel juridisk diktering, vil materialet måtte utvides med nye transkripsjoner.

I den forbindelse er det grunn til å understreke at det er utviklet et omfattende sett av leksikalske verktøy, for det meste skrevet i Perl, som forenkler og automatiserer arbeidet med leksikalske ressurser. Disse verktøyene, som blant annet omfatter verktøy for grafem-til-fonem-konvertering, vil kunne være til hjelp ved fremtidige utvidelser av leksikonet.

Valget av Ålesund bymål som grunnlagsdialekt for det norske leksikonet kan synes for spesifikt. Det norske leksikonet danner likevel grunnlag for mer allmenne taleteknologiske applikasjoner, blant annet den nye norske IBM-baserte talesyntesen, som er basert på Oslo-dialekten (stemmen kalt Henrik).

Leksikonene er kvalitetssikret ved interne rutiner på en slik måte at formatet er entydig, transkripsjonene kun inneholder lovlige tegn og tegnkombinasjoner, og den morfosyntaktiske annoteringen kun inneholder lovlig merking. Det er imidlertid ikke dokumentert i hvilken grad det er konsistens i transkripsjonene. Det er på det rene at konvensjonene har endret seg underveis. Enkeltstående eksempler på mangel på konsistens er oppdaget, for eksempel at et ord som *Årdal+s+veien* /"o:$d`A:ls$%v{*I$@n/ er dekomponert, mens *Årdalstangen* /"o:$d`A:l$%stAN$@n/ ikke er det. Dette er en konsekvens av innføring av en regel om å dekomponere egennavn, og den endrete praksisen fører til inkonsistens med hensyn til stavelsesinndeling ved fuge-s. Hvorvidt eksisterende transkripsjoner er endret som følge av denne og andre konvensjonsendringer er ikke dokumentert.

Standarden for leksikalske databaser er beskrevet i ELRA-dokumentet http://www.spex.nl/validationcentre/d11v21.doc, avsnitt 3.6. og i http://www.elra.info/services/validation_manual_lexica.pdf. Det er på det rene at NSTs leksikalske databaser oppfyller de formelle kravene til format, representasjon, dokumentasjon og samsvar med akustisk base som denne standarden stiller. Det må likevel påpekes at det ikke finnes noen dokumentert dekningsgrad for leksikonene. For allmennspråklige leksikon gis følgende spesifikasjon fra ELRA: ”In a general language lexicon the closed classes (e.g. pronouns, determiners, articles and prepositions) and series (e.g. auxiliary verbs, modal verbs, days of the week, months of the year) are expected to have 100% coverage. The open classes (nouns, verbs, adjectives etc.) are expected to be represented with a frequency reflecting their relative frequency in the language.” Det er ikke funnet noen dokumentasjon som eksplisitt angir denne dekningsgraden i NSTs materiale, men det faktum at f.eks. det norske leksikonet blant annet omfatter ordtilfanget fra Bokmålsordboka skulle tilsi en høy generell dekningsgrad.

Leksikonet baserer seg på internasjonalt anerkjente formater som Parole/SIMPLE-formatet for morfosyntaktisk annotering og SAMPA-standarden for fonetisk transkripsjon.

Samlet sett kan det konkluderes med at NSTs leksikalske databaser utgjør en verdifull ressurs som er i all hovedsak er i samsvar med internasjonal standard for slike språkressurser.

3.4 Korpus

NSTs tekstkorpus ligger i sin helhet på \\Server2. Katalogen \\Server2\KorpusOrginal inneholder originalversjoner av tekstleveranser for norsk, svensk og dansk. Katalogen \\Server2\KorpusWork inneholder det som finnes av bearbeidelser av tekstmaterialet. Dette omfatter både annoterte korpus og språkmodeller basert på materialet.

Tekstene er samlet inn dels gjennom henvendelser til eksterne tekstleverandører som Aftenposten og Aschehoug, og dels gjennom manuell eller automatisk nedlasting av internett-tekst. Det er generelt tilfelle at tekstene er samlet inn med det formål å utvikle statistiske modeller basert på materialet, samt som utgangspunkt for frekvensbaserte leksikalske databaser. Tekstene er også benyttet til å lage fonetisk balanserte manuskripter for opplesning til akustisk database, og som grunnlag for å produsere innlesningsmanuskript for talesyntese.

Det er begrensninger i bruken av materialet for en rekke av tekstene. Opphavsrett/ eierskap til materialet ligger generelt hos tekstleverandør. Det er også begrensninger for å overlate bruken til andre. Dersom materialet skal inngå i noen annen sammenheng er det dermed et behov for å reforhandle avtaler i en rekke tilfeller. Generelt er tekstene forbundet med en standardavtale som ofte har følgende ordlyd:

> *NST får rett til å bearbeide dette elektronisk for:*
> - *Utarbeidelse av lister over de hyppigste ordene, frasene, forkortingene, egennavnene, akronymene etc. i norske tekster*
> - *Grunnlag til opplesing ved digitale opptak av norskspråklige fra forskjellige dialektområder*
> - *Grunnlag til statistiske modeller, tekstanalyse, talegjenkjenning, diktering og talesyntese*
> - *Oppbygging av NSTs tekstkorpus*
>
> *NST kan ikke overdra retten til å benytte materialet til andre, i denne eller annen sammenheng. Deler av arbeidet vil foregå ved NSTs belgiske samarbeidpartner Lernout & Hauspie, som vil få tilgang til materialet gjennom denne avtale.*

I det følgende gis en tabellarisk oppstilling av NSTs tekstkorpus for norsk, svensk og dansk. Enkelte av tallene for antall ord er estimater. For svensk og dansk har det ikke vært kapasitet til å i undersøke hver enkelt tekstleveranse. Beskrivelsen av tekstformater er derfor noe mer detaljert i det norske korpuset.

3.4.1 Norsk

3.4.1.1 Oversikt over bokmålskorpus

Tabellen viser NSTs norske korpus av tekster i ubearbeidet versjon.

Navn/ leverandør	Innhold/sjanger	Fil-typer	Formatbeskrivelse	Årg.	Ant. filer	Str. MB	Ant. ord
Aftenposten	Avistekst: nyheter, sport, underholdning, feature, etc.	.txt	Tekstfiler med hele avisartikler og utfyllende metainformasjon inkl unik artikkel-nummerering, dato osv. Ikke annotert.	1984-1998	390	3040	344 000 000
Arbeiderpartiet i Bergen	Saksdokumenter, årsmøterapporter.	.doc	Word-filer med formatering, tabeller etc.	1996-2001	15	4	1 600 000
Aschehoug	Romaner, øvrig skjønnlitt., faktalitteratur, lærebøker. Svært variert materiale, delvis oversatt litt. I overkant av 450 bøker.	.txt .pdf .doc .qxd .ps .w51 .rtf	Bøker i fulltekstversjon, én bok per fil. En rekke ulike formater, inkl. proprietære (WordPerfect, QuarkXpress), en god del MS Word. Noe grafikk. Mangelfull dokumentasjon. Behov for konvertering og standardisering. Ett ukjent filformat (”tekst”).	1990-1999	3425 +2574	1490 +1360	6 900 000
Bergens Tidende	Avistekst: nyheter, sport, underholdning, feature, etc.	Rdf	Tekstfiler med hele avisartikler og utfyllende metainformasjon inkl unik artikkel-nummerering, dato osv. Kodet med xml-liknende tagger.	1990-1999	111	947	80 900 000
Korrespondanse	Bedriftsrelatert kommunikasjon, brev, notater.	.doc	Bifogatfrånemailkampanjen.zip Word-dokumenter, delvis med brevhodegrafikk	1998-2000 (?)	30	6	50 000
Computerworld	Nyhetstekst, IT	.txt	Tekstfiler med hele avisartikler og utfyllende metainformasjon inkl unik artikkel-nummerering, dato osv. Ikke kodet.	1993-1999	6	83,4	12 100 000
Elektronikkforlaget	Nyhetstekst, IT	.txt	Korte tekstfiler uten metainformasjon.	Før 2000	364	1,6	1 500 000
FinansFokus2000	Nyhetsmeldinger fra Finansforbundet	.doc	Word-dokumenter uten metainformasjon.	Før 2000	173	4,5	150 000
Høyre	Politiske dokumenter, korrespondanse, taler, program, notat	.doc	Hovedsakelig word-dokumenter, også et ukjent format. Lite metainformasjon.	1996-2000	4536		1 600 000
IT-Avisen	Nyhetstekst, IT	.txt	Tekstfiler med noe metainformasjon. Ikke kodet. Korte filer med én nyhetsmelding per fil.	1996-2000		37,1	3 850 000
Lerum	Romaner av May Grethe Lerum, alle bøkene i serien Livets Døtre	.txt	Tekstfiler med en roman per fil.	1990-tallet	36	10,5	1 970 000
Magma	Nyhetsmagasin, fagartikler, økonomi.	.html	Kodet i html, noe metainformasjon.	1998-2000	321	7,7	700 000
Morgenbladet	Nyhetstekst, generell	.doc	Word-dokumenter med tekstformatering men uten metainformasjon. Én artikkel per fil.	Før 2000	47	0,902	50 000
Nærings- og handelsdep.	Offentlig korrespondanse	print	Én enkelt tekstfil med brev, inkl. metainformasjon som saksnummer, dato osv.	1989-1992	1	64,1	6 900 000
Nærlingslivets ukeblad	Uleselig	.html	Uleselig		1	5,0	
Nettavisen	Nyhetstekst, generell	.DA0 .DA1 .DAT	Utdrag fra SQL-database, tre enkeltfiler med mange nyhetsartikler.	Før 2000	964	2380	25 000 000
Nordisk råd	Protokoller og annen dokumentasjon	.html	Kodet i html, noe metainformasjon.	Før 2000	9	0,1	
Nordisk råd press	Pressemeldinger fra Nordisk råd/Nordisk ministerråd	.html	Kodet i html, noe metainformasjon, en pressemelding per fil.	1997-1999	58	0,2	
Nordland Fylke	Offentlig korrespondanse	.doc	Word-dokumenter med brevhodegrafikk. Finnes også som tiff-filer.	1998	2736	167	380 000
NTB	Nyhetstekst, generell	.txt	Tekstfiler med hele avisartikler og utfyllende metainformasjon inkl unik artikkel-nummerering, dato osv. Ikke annotert. (Samme format som Aftenposten.)	1985-1998	419	2240	286 000 000

Navn/ leverandør	Innhold/sjanger	Fil-typer	Formatbeskrivelse	Årg.	Ant. filer	Str. MB	Ant. ord
Revisorhandboken	Revisors Håndbok 2000 (21. utgave)	.txt	Én enkeltstående tekstfil uten annoteringer eller metainformasjon. Finnes pgså som pdf.	2000	1	4646	677 000
Sandemo	Romaner i serie (Isfolket etc.)	.txt	Én tekstfil per roman, uten annoteringer eller metainformasjon.	Før 2000	14	4,24	790 000
Statsbudsjettet	Off. dok., Statsbudsjettet Ot. prp. nr. 1 (2001)	.txt	Tekstfiler uten annoteringer eller metainformasjon.	2001	4	3,0	425 000
Teknisk ukeblad	Nyhetsmeldinger, teknisk domene	.html	Et stort antall filer med indeks og enkeltartikler. Kodet i html men uten metainformasjon.	1996-2000	11744	165	1 454 000
Tekstinnsaml_bokmaal	Internett-tekst fra en rekke ulike domener	.txt	Nedlastete tekster, konvertert til rent tekstformat. Én nettside per tekstfil, uten annoteringer eller metainformasjon.	Før 2000	110948	1130	125 100 000
TV2	Nyhetstekst, generell	.mdb	Microsoft Access Database, uten annoteringer, noe metainformasjon.	Før 2000	1	184	8 350 000
Usenet	Diskusjonslister, e-post, usenet fra en rekke ulike domener	.txt	Nedlastete tekster, konvertert til rent tekstformat. Én debattliste per tekstfil, uten annoteringer elle metainformasjon.	Før 2000	398	1361	4 200 000
	Nyhetsgrupper, Telenor	.txt .dbx	Tekstfiler og proproetært format (Outlook Express). Én debattliste per tekstfil, minimalt med koding eller metainformasjon.	Før 2001	78	14,9	10 000 000
Ustad	Romaner i serie, (Fire søsken)	.txt	Én roman per tekstfil. Ingen koding eller metainformasjon.	Før 1999	3	1,01	190 000
Venstre	Politiske dokumenter, korrespondanse, taler, program, notat	.html .rtf	Én tekst per fil, hovedsaklig kodet i html, ikke metainformasjon.	Før 2001	1655	13591	1 600 000
Nytt internett	Site, finance	.html	Én tekst per fil, hovedsaklig kodet i html, ikke metainformasjon.	Før 2001			25 000 000
Site, IT	Site, IT	.html	Én tekst per fil, hovedsaklig kodet i html, ikke metainformasjon.	tmengder. Før 2001			23 346 000

Materialet utgjør ca. 975 millioner ord løpende råtekst. Dette er for det meste bokmålstekst.

3.4.1.2 Oversikt over nynorskkorpus

NST har ikke utviklet produkter for nynorsk, og informasjonen om dette materialet er langt mindre detaljert. Nynorskmaterialet består av følgende datasett.

Navn/ leverandør	Innhold/sjanger	Fil-typer	Formatbeskrivelse	Årg.	Ant. filer	Str. MB	Ant. ord
gamalt_fraa_voss	ukjent	ukjent	40 filer, ukjent binært filformat.	1995	40	0,77	
Internet_KK	Internettekst fra 4 kilder: Førde kommune, Hordaland fylkeskommune, Sogn Avis, Firda.	.html	Nedlastet internettekst med grafikk, kodet i html. Svært mange små filer.	2001			
nn_fra_bm_internet	For det meste nyhetstekst, overført	.txt	Rene tekstfiler, for det meste uten koding	1998	2957	150	

	fra bokmålskorpuset.		eller metainformasjon.				
Tekstinnsaml_nynorsk	Internett-tekst fra en rekke ulike domener	.txt	Nedlastete tekster, konvertert til rent tekstformat. Én side per tekstfil, uten annoteringer eller metainformasjon.	1998	5870	60	

3.4.1.3 Medisinsk tekst

NST har samlet inn medisinsk tekst for norsk og svensk. Materialet er ordnet etter medisinsk fagområde, og det norske originalmaterialet befinner seg i følgende kataloger:

Kardiologi: \\Server2\KorpusMedisin\no_kar_korpus\original_korpus
Patologi: \\Server2\KorpusMedisin\no_pat_korpus\original_korpus
Radiologi: \\Server2\KorpusMedisin\no_rad_korpus\original_korpus

Det medisinske korpuset består i sin helhet av rene tekstfiler. Disse er store, og formatet er gjennomgående én artikkel per linje. Materialet fremstår som udokumentert og noe kaotisk. Det har vært vanskelig å danne seg et bilde av dets størrelse og grad av bearbeiding, men det er på det rene at tekstene er anvendt for å utvikle IBM-baserte språkmodeller til bruk i programvaren Autor, som er NSTs program for medisinsk diktering.

Det later til at kardiologidelen ikke er bearbeidet, mens de øvrige delene har i ulik grad blitt bearbeidet. Patologitekstene har kun vært gjenstand for basal tekstrensing og ikke øvrig bearbeiding, mens radiologidelen har blitt bearbeidet i høy grad i forbindelse med utvikling av medisinsk diktering for radiologi. Bearbeidingen har omfattet nødvendig anonymisering, tekstnormalisering, stavekontroll og tokenisering på setningsnivå. Filen \\Server2\KorpusMedisin\no_rad_korpus\whole_corpus.txt inneholder hele radiologikorpuset som én tekstfil.

Det medisinske korpuset har følgende omfang (unntatt kardiologi):

Patologi	original	renset
Trondheim	11 746 439	10 369 360
Ullevål	19 127 739	17 124 132
Sum	30 874 178	**27 493 492**

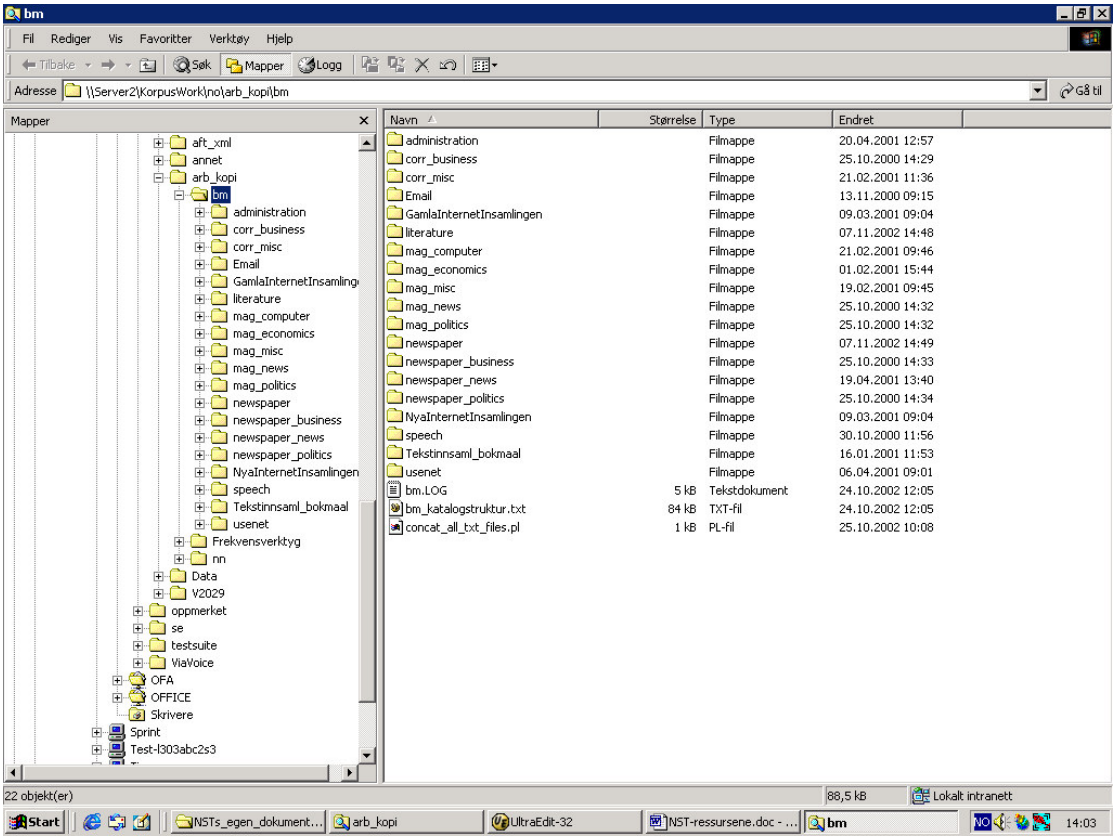
Radiologi	original	renset
Oslo	16 705 576	16 698 126
Tønsberg	19 106 408	19 281 249
Tromsø	39 150 890	37 997 825
Trondheim	16 712 156	13 407 866
Sum	91 675 030	**87 385 066**

Tallene er hentet fra filen \\Server2\KorpusMedisin\wordcounts_all_medical.xls.

3.4.1.4 Innhold og grad av bearbeiding

Det generelle norske tekstkorpuset er i noen grad bearbeidet. En bearbeiding har vært nødvendig for produksjon av manuskriptsetninger og frekvensbaserte leksikalske data (blant annet en såkalt 100k-liste, bestående av de 100 000 mest frekvente ordformer).

Det bearbeidete norske korpuset finnes på \\Server2\KorpusWork\no\arb_kopi\bm. Mens råtekstene er organisert etter leverandør, er det bearbeidete korpuset organisert etter sjanger, som vist i figuren:



Informasjon om frekvensen av ordformer finnes i katalogen \\Server2\KorpusWork\no\arb_kopi\Frekvensverktyg og en oversikt over korpusets katalogstruktur og enkeltkatalogers størrelse finnes på \\Server2\KorpusWork\arb_kopi\no\bm_katalogstruktur.txt . Merk at en del av katalogene er tomme, nemlig

corr_business
mag_news
mag_politics
newspaper_business
newspaper_politics

Det bearbeidete korpuset består av rene tekstfiler som utgjør anslagsvis 735 millioner ord løpende tekst, mens korpuset av rådata består av 975 millioner ord, som vist i tabellen i forrige avsnitt.

Selve bearbeidingen omfatter konvertering fra proprietære formater til rene tekstfiler, fjerning av duplikate tekster, rensing og forkasting av ukurante filformater (tiff-filer fra skannete dokumenter, QuarkExpress, FrameMaker osv.). Sistnevnte tiltak reduserer materialets størrelse en del, men ikke dramatisk. For eksempel opplyses det i en loggfil at for tekstene fra Aschehoug at "disse utgjør store mengder i MB men

ikke i antall ord (bare 5 av totalt 50 bøker)". Verktøy som er anvendt ved bearbeiding av tekst finnes på \\Server2\KorpusWork\Felles\VERKTYG.

I likhet med mange andre tekstinnsamlingsprosjekter har man satt sammen et materiale basert på det man kan få tak i, og ikke ut fra et mål om en bestemt fordeling av tekster i henhold til sjanger og teksttype. Avistekst er åpenbart overrepresentert, mens korrespondanse (særlig e-post), skjønnlitteratur og faglitteratur er underrepresentert. Fordelingen på tekstkategori er som følger:

|- Email (820,47 Kb)	0,01 %
|- GamlaInternetInsamlingen (1,46 Gb)	22,89 %
|- NyaInternetInsamlingen (0 Byte)	0,00 %
|- Tekstinnsaml_bokmaal (783,27 Kb)	0,01 %
|- administration (54,03 Mb)	0,85 %
|- corr_business (0 Byte)	0,00 %
|- corr_misc (66,88 Mb)	1,03 %
|- literature (57,60 Mb)	0,89 %
|- mag_computer (101,55 Mb)	1,58 %
|- mag_economics (983,80 Kb)	0,02 %
|- mag_misc (4,62 Mb)	0,07 %
|- mag_news (0 Byte)	0,00 %
|- mag_politics (0 Byte)	0,00 %
|- newspaper (3,47 Gb)	54,41 %
|- newspaper_business (0 Byte)	0,00 %
|- newspaper_news (1,15 Gb)	18,03 %
|- newspaper_politics (0 Byte)	0,00 %
|- speech (12,57 Mb)	0,19 %

Materialet kan dermed ikke hevdes å utgjøre et balansert allmennkorpus. En fordeling etter antall ord basert på råtekstene finnes på \\Server2\KorpusWork\Felles\Documentation\ressursoversikter\ferskeste tall NO DK SE.xls

Bearbeidingen begrenser seg til en ren konvertering til tekstformat. Tekstene gjengir all tekst som forekommer i grunnlagsdokumentene, som vist i teksteksempelet.

Teksteksempel: Nordland Fylkeskommune (konvertert fra Word)

```
NORDLAND FYLKESKOMMUNE  Vår dato:   Vår referanse:    Arkivnr:
Fylkesrevisjonen  08.09.1998  98/04171-1
      Vår referanse må oppgis ved alle henvendelser

      Deres dato: Deres referanse:

Saksbehandler:  ****,  Tlf. 75 50 04 71

**** ****
****

8215 VALNESFJORD
```

```
REVISJONSRAPPORT NR. 1 1998 -  STIFTELSEN ****

Vi er nå ferdige med gjennomgangen av stiftelsen ****'s regnskap pr.
31.03.98. Som avtalt
oversendes vår skriftlige oppsummering av revisjonsarbeidet.

Den økonomiske situasjonen
Av regnskapet fremgår det at den økonomiske situasjonen ikke er
forbedret i løpet av årets tre
første måneder. Våre merknader i revisjonsberetningen for 1997 samt
revisjonsrapport nr 2/97 er
dermed ennå aktuelle. Vi går ut fra at dere følger denne utviklingen
nøye og holder oss orientert.

Husleieinntekter
Ut fra fremlagt prisliste av 01.08.96 er det ved vår kontroll
registrert at det er benyttet feil sats for
en del boliger i ****. I hovedsak har de ansatte betalt lavere
månedsleie enn hva prislisten
tilsier. Etter hva vi forstår er dette basert på et styrevedtak, vi
ber om nærmere opplysninger.

Boligene dette gjelder er:
Adresse     Type        Prisliste   Regnskap    Leietaker
****. 7     hovd.lei    3.400       2.000       **** **** (**** ****
til
                                                15.02.98 betalt 1.700
)
    "   10      "       2.800       2.200       **** ****
    "        13       "       2.800       2.000       **** ****
    "        13   sokkel            1.500       1.700       **** ****

Med hilsen

**** ****
**** ****
```

Anonymisering av korrespondanse er ikke foretatt, noe som vil være påkrevd ved allmenn bruk av materialet. I teksteksemplet indikerer symbolet **** anonymisering foretatt av rapportens forfatter. Det forekommer tabeller og lister som del av korpustekstene uten at dette er eksplisitt markert.

Materialet mangler enhver form for oppmerking (tagging) av lingvistisk informasjon (ordklasse, lemma). Riktignok er deler av grunnlagsmaterialet strukturelt oppmerket i avsnitt, overskrifter, bildetekst og brødtekster, men dette skyldes a priori oppmerking fra tekstleverandør, og følger ikke noen enhetlig standard definert innefor selve korpusprosjektet, og heller ikke eksterne xml/sgml-standarder for tekstkoding.

Eksempel på leverandørkodet tekst (BT)

```
<artikkel>
<BRS Assigned Accession Number>000007586
<Merknad>BYSIDEN
<Dato>19901020
<Dokumentid>5625
<ILLU>2 S/H
<Side>14
```

```
<SPR>BM
<STOR>1725
<EMNE>
     BOKBRANSJE; FORFATTER; ARRANGEMENT; FORHÅNDSOMTALE
<Tittel>VAKENATT MED LESEFEST
<Brødtekst>
     Vakenatt med lesefest

     - Det stundar til bokdag, eit tradisjonelt haustarrangement for
     litterært interesserte menneske. I år snur dei litt på flisa og
gjer
     bokdagen til kveld, ein lang lang kveld som først endar ein time
over
     midnatt.

     Forfattarane vil gjera dette til ei "vakenatt for biblioteka""
til
     støtte for biblioteka og ein honnør for det dei gjer for å auka
leselyst
     mellom folk. Dette blir understreka førstkomande laurdag i
Schøtstuene
     der Finn Bjørn kåserer om "Gledens hus - bilioteket". Lesefesten
er då
     alt opna med fanfare og Harald Sverdrup som les eigne dikt.

     ...

Willy Dahl innleier til samtale om far- og barn.
</artikkel>
```

I de tilfeller hvor det har fantes oppmerking i grunnlagsmaterialet er denne ivaretatt i de konverterte tekstfilene, som vist i eksemplet, men generelt mangler det strukturell merking. For eksempel inneholder en skjønnlitterær tekstfil både innledende informasjon om trykking, ISBN-nummer, forfatterens bibliografi osv. i tillegg til selve kjerneteksten, uten at dette er eksplisitt markert.

Interpretativ koding er fraværende, lemmatisering, merking av setningsgrenser og lignende er ikke foretatt. Korpuset er heller ikke tilgjengelig i et enhetlig grensesnitt.

NST later imidlertid til å ha startet en prosess med xml-basert oppmerking av korpuset. På \\Server2\KorpusWork\oppmerket\no\bm\ITAvisen finnes tekster som er merket med strukturell informasjon.

Eksempel på NST-kodet tekst

```
<artikkel>
<tittel>D-dag for DT</tittel>
<undertittel></undertittel>
<forfatter>Øystein Kvistad</forfatter>
<datotid>1996-10-22 00:00:00</datotid>
<ingress>Deutsche Telecom vil idag presentere et prisframlegg på
mellom 25 og 30 tyske mark for de 500 nye aksjene som skal selges i
det største offentlige tilbud som noen sinne er blitt presentert i
Europa.</ingress>
<tekst>Dette priser Deutsche Telecom til mellom 65 og 78 milliarder
D-mark, noe som gjør selskapet til et av de mest verdifulle i Europa.
Financial Times skriver i dag at 70 prosent av de nye aksjene vil bli
solgt på det tyske markedet - det vil si omlag ti prosent mer enn det
som ble antydet i det opprinnelige prospektet.Det er ventet at
```

```
Deutsche Telecom også vil plassere 15 prosent av aksjene på det
amerikanske markedet og ti prosent til institusjonelle investorer i
England.
</tekst>
</artikkel>
```

Mengden av tekst som har vært gjenstand for slik bearbeiding er svært liten, og omfatter kun 3,8 millioner ord (0,4 prosent av hele materialet), og kun tekster fra IT-avisen. Det finnes ikke noe tilsvarende nynorskmateriale.

Det må understrekes at den type oppmerking som ELRA-dokumentet omtaler ikke alltid er en forutsetning for bruk; for visse formål kan det være tilstrekkelig å benytte et korpus uten ordklassetagging, for eksempel.

Deler av det norske tekstmaterialet har vært brukt som grunnlag for stokastisk modellering og tekstbasert n-gramproduksjon. Dette dreier seg om språkmodeller som er utviklet ved hjelp av IBM-teknologi. Disse finnes på følgende steder:

\\Server2\KorpusWork\cno13
\\Server2\KorpusWork\modeller\cno6
\\Server2\KorpusWork\modeller\cno7

og lister over n-gram finnes på \\Server2\KorpusWork\ngram_lister.

3.4.2 Svensk

Det svenske korpuset har gjennomgått en bearbeiding som er mer omfattende enn det norske. Dette er gjort i forbindelse med leveranser til L&H for utvikling av et svensk dikteringssystem.

3.4.2.1 Oversikt over svensk korpus

Tabellen viser NSTs svenske tekstkorpus av tekster i ubearbeidet versjon. Det har ikke vært tid nok til å gjennomgå det svenske tekstkorpuset i detalj, og ressursoversikten er derfor noe mindre detaljert med hensyn til formatbeskrivelser enn den tilsvarende norske.

Navn/ leverandør	Innhold/sjanger	Format og filtyper	Årgang	Ant. filer	Str. MB	Ant. ord
Affärsvärlden	Næringslivsnyheter	Txt	før 2000	6	113	9 849 708
Affärsdata	Ca. 15 ulike tidskrifter	Txt	før 2000	2038	2,9 GB	80 000 000
Aftonbladet	Dagsavis	Txt	før 2000	20	201	30 905 384
Offentlig förvaltning	Protokoll, rapporter, korrespondanse	Txt, doc, pdf	før 2000	>25 000	>2 GB	6 500 000
Datateknik3.0	Datatidsskrift	Txt, xls	før 2000	2	17,2	2 300 000
Email	Korrespondanse	Doc, msg, eml, txt	før 2000	4596	73,8	1 000 000
Forskning och framsteg	Teknisk journal	doc, qxd	før 2000	512	628	1 300 000
Gbgtidn	Dagsavis	Txt	før 2000	1	32,6	5 259 431
Ica-Kuriren	Dameblad	Quark Express	før 2000	5842	2,76 GB	2 200 000
LexiLogik	Avisartikler	Txt	før 2000	7	288	39 653 768
LexiLogik	Romaner	txt	før 2000	30	18,7	3 221 796
Internet	Diverse	Html, shtml, htm, txt, pdf, doc	før 2000	>100 000	>1500	68 000 000
Ny teknik	Teknisk tidsskrift	Txt	før 2000	1	41,6	3 406 932
Presstext	Avisartikler	Txt Tidsbegrenset, utlån, fra 1/11-1999	1990-1999	10	1790	297 656 290

		til 31/10-2001				
Samhall	Romaner	Txt	før 2000	98	211	37 396 004
SvD	Artikler fra internett	Txt	før 2000	104	1,41	
Sveriges television	TV-manuskripter	Txt	før 2000	242	6,79	4 000 000
Internetworld	Datatidsskrift	Txt	før 2000	5708	156	600 000

Tabellen angir det som finnes av ubearbeidet tekst. Samlet utgjør det generelle korpuset ca. 593 millioner ord.

3.4.2.2 Grad av bearbeiding

Det svenske korpuset er noe mer bearbeidet enn det norske. Dette kan tilskrives leveranser til L&H i forbindelse med utvikling av svenskspråklig versjon av L&Hs dikteringssystem.

Bearbeidete svenske korpusdata ligger på følgende områdene

\\Server2\KorpusWork\se\
\\Server2\KorpusWork\Leveranser_svensk_korpus

Det har vært vanskelig å danne seg et bilde over hvordan rensingen har foregått, og hva som er seneste versjon av den bearbeidete teksten. Imidlertid antas det at leveransene omfatter det fullstendige bearbeidete materialet i sin reneste form. Dette materialet er anslagsvis på 100 millioner ord.

3.4.2.3 Juridisk tekst

NST har samlet inn svensk juridisk tekst fra ca 90 ulike leverandører. Denne finnes på \\Server2\KorpusWork\Legal. Denne delen av korpuset fremstår som udokumentert, men ifølge en ressursoversikt består det av 44 600 000 ord. Materialet består av tekstfiler og html-filer. Det ser ikke ut til å være bearbeidet i noen grad.

3.4.3 Dansk

Det har ikke vært tid nok til å gjennomgå det danske tekstkorpuset i detalj, og oversikten nedenfor er derfor begrenset til en overordnet kvantitativ beskrivelse.

Navn/ leverandør	Innhold/sjanger	Format og filtyper	Årgang	Ant. filer	Str. MB	Ant. ord
BT, Berlingske avisdata		txt, dat	1995-1999, 2000??	127	1900	146 000 000
Ekstrabladet						56 186 364
Politiken		txt				117 448 174
TekniskForlag		doc		51558	502	20 678 866
HjerneSagen	articles	doc		833	12,7	247 122
Internet		txt, htm, html, shtml, doc		35085	1020	58 105 125
korrespondanse						1 331
email	ltters	txt, doc		5	0,294	564 391
usenet						3 230 720
Lev	articles	doc		370	7,75	237 247
økologisk raad	articles	html, rtf		253	5,29	178 855
romaner		doc		2	0,919	100 208
Annet	thesis	dic, txt		20	26	4160177

Samlet utgjør materialet ca. 407 millioner løpende ord. Det ser ikke ut for å være samlet inn dansk tekst innenfor spesialdomener som medisin eller jus.

3.4.4 Kvalitetsvurdering

NSTs tekstmaterialet befinner seg i originalversjon på \\Server2\KorpusOrginal\ i form av komprimerte filer. Hver post i tabellen tilsvarer én zip-fil. Beregningene av størrelse er til dels basert på tidligere dokumentasjon (041212_OversiktSpråkressurser_en.doc).

Som det fremgår av oversikten er NSTs korpus meget omfangsrikt, og det fremstår som svært variert og innholdsrikt, og representerer et svært bredt omfang av tekstkategorier og bruksområder. For samtlige språk omfatter materialet i seg selv tilstrekkelige tekstmengder i henhold til det som kreves av et tidsriktig allmennkorpus.

Likevel må det understrekes at en fremtidig bruk av NSTs norske korpus reiser atskillige utfordringer som angår

- kompleksiteten i filtyper
- mangelen på enhetlige tekstformater
- mangelen på tekstkoding
- ubalanse bokmål/nynorsk
- fordeling av teksttype
- materialets årgang
- juridiske bindinger

Utfordringene knyttet til det første punktet har NST selv løst ved at materialet er konvertert til rene tekstformater, uten betydelig reduksjon av tekstmengde.

Det andre punktet innebærer at det kreves atskillig arbeid for å få materialet til å fremstå som et helhetlig korpus. Dette vil kunne være ressurskrevende, særlig i de tilfeller hvor det dreier seg om postscriptfiler og proprietære/lukkede formater.

Internasjonal standard for korpus er beskrevet i ELRA-dokumentet *Validation of Linguistic Corpora* (http://www.elra.info/services/valid/wp3/index.htm), som baserer seg på standarden EAGLES Guidelines for morpho-syntactic annotation (Leech & Wilson 1994). ELRA-dokumentet lister følgende kriterier for validering av tekstkorpus:

- fins tagging?
- er taggingen korrekt?
- er taggingen konsistent?
- er taggingen i samsvar med eksternt definert standard?
- er taggingen fullstendig i henhold til eksterne krav til obligatoriske trekk?

Beskrivelsen i de språkspesifikke avsnittene ovenfor viser at ingen av disse kriteriene kan sies å være oppfylt i NSTs korpus. En stor del av materialet består av tekstfiler uten annoteringer eller metainformasjon. En standardisering og tydelig merking av hva som er selve teksten og hva som er metainformasjon er nødvendig for enhver teknologisk eller forskningsmessig bruk, enten det dreier seg om stokastisk modellering, tekstbasert n-gramproduksjon, frekvensbasert generering av leksikalske data eller språkvitenskapelige studier. Mangel på slik koding vil føre til feile opplysninger om bruksfrekvens og kollokasjoner.

I tillegg er det problematisk at materialet hovedsakelig er samlet inn i en periode frem mot årtusenskiftet, og at det således stadig blir mer utdatert. En konsekvens av dette vil være at materialet har lav dekningsgrad av neologismer og importord av nyere dato, og at frekvensopplysninger ikke er oppdaterte. For Aftenpostens del er ca. halvparten av materialet fra før avisens språkreform, hvilket vil si at det inneholder riksmål. NST måtte selv forholde seg til en femårsregel ved leveranser til L&H, dvs. at tekster eldre enn fem år ved leveransedato ble betraktet som utdaterte.

Et ytterligere poeng, som kan være et problem i visse sammenhenger, er at det er stor ubalanse mellom bokmål og nynorsk. Fordelingen på målformene er basert på filstruktur og ikke på analyse av selve tekstene. Det er derfor grunn til å tro at det finnes mer nynorskmateriale enn det som tabellene viser, fordi bl.a. Bergens tidende inneholder en god del nynorsktekst.

Med disse innvendingene er det imidlertid ikke ment å antyde materialet ikke kan ha et betydelig brukspotensial både som språkteknologisk grunnlagsressurs og som forskningsmessig studieobjekt. Tekstinnsamling er i seg selv en ressurskrevende aktivitet, og det er grunn til å fremheve arbeidet som er utført med å fremskaffe et bredt sammensatt og omfattende materiale. Imidlertid krever enhver bruk både videre bearbeiding av materialet og avklaring av juridiske forhold, overføring av bruksretter, reforhandling av avtaler osv.

4 Kort sammenfatning

I denne rapporten er det foretatt en gjennomgang av NSTs ressurser på Voss. Det har vært lagt vekt på å dokumentere hva som finnes av språklige primærressurser og ressursenes plassering, omfang, format og kvalitet. I gjennomgangen har det vært skilt mellom følgende hovedkategorier:

- akustiske databaser for talegjenkjenning
- akustiske databaser for talesyntese
- leksikalske databaser
- tekstkorpus

For hver enkelt kategori er det vurdert om ressursene samsvarer med internasjonale standarder for språkressurser, slik de er definert av organisasjonen ELRA. De spesifikke konklusjonene for dette er gitt i avsnittene 3.1.8, 3.2.4, 3.3.6 og 3.4.4. Det er kun NSTs egne oppbygde ressurser som har vært gjenstand for en slik evaluering.

Den desidert største og mest verdifulle ressursen er NST egenutviklete akustiske database for talegjenkjenning. Rapporten fremhever at lydmaterialet er av generell høy kvalitet, grundig validert, og produsert i henhold til gjeldende standarder for taledata. Det foreligger avtaleskjema fra hver enkelt informant i NSTs arkiv, og disse gir rett til å beholde og nytte seg av opptakene til taleteknologiske formål. Det anbefales at NSTs akustiske database ivaretas og gjøres tilgjengelig for fremtidige forsknings- og teknologiutviklingsformål.

Talesynteseopptakene fordeler seg på tre kategorier: a) opptak gjort til programvaren RealSpeak i Belgia, b) opptak gjort i kontormiljø for testformål, og c) opptak gjort i studio på Voss for reell produktutvikling. Kategori a) er uaktuelle for videre bruk da de ikke befinner seg i noe lagringsformat på Voss, fordi produktutviklingen i sin

helhet foregikk i Belgia. Kategori b) omfatter hovedsakelig NST-ansatte som ble spilt inn for ulike testformål, og opptakene har svært liten gjenbruksverdi. Kategori c) er verdifulle høykvalitetsopptak med svært god annotering. Til disse opptakene er det rekruttert profesjonelle skuespillere. Det foreligger avtaler med hver av disse i arkivet på Voss. Det anbefales at dette materialet ivaretas og gjøres tilgjengelig for fremtidige forsknings- og teknologiutviklingsformål.

NSTs norske, svenske og danske leksikalske databaser er, i likhet med den akustiske databasen, en egenutviklet ressurs hvor det ikke knytter seg avtalemessige problemstillinger. I rapporten er det reist visse spørsmål knyttet til manglende tilførsel av nytt vokabular de seneste år, grad av konsistens, og manglende dokumentasjon av dekningsgrad. Likevel må det understrekes at disse ressursene er omfattende og at det har krevd betydelig manuelt arbeid å opparbeide dem. Leksikonene baserer seg på internasjonalt anerkjente formater som Parole/SIMPLE-formatet for morfosyntaktisk annotering og SAMPA-standarden for fonetisk transkripsjon. De er i all hovedsak i tråd med gjeldende internasjonal standard. De leksikalske databasene er en nødvendig ressurs for å kunne fullt ut nyttiggjøre seg den akustiske databasen, og det anbefales at denne også gjøres generelt tilgjengelig. Det er også grunn til å fremheve de mange egenproduserte språkbehandlingsverktøy som forenkler og automatiserer arbeidet med leksikalske ressurser. Disse vil være et verdifullt supplement til de leksikalske ressursene.

Korpusressursene er svært omfattende. Her er det imidlertid en del uavklarte juridiske spørsmål knyttet til bruksrett (jf. Breiviks rapport). Det avtalemessige er problematisk fordi korpuset er bygget opp ved hjelp av eksterne leveranser fra en lang rekke skandinaviske kilder. Det har også vært pekt på betydelige teknologiske problemstillinger knyttet til bruk av korpuset, først og fremst at størstedelen av materialet er utdatert (hovedsaklig fra før 2000), dernest at det ikke fremstår som et helhetlig, kodet tekstkorpus men som en stor og heterogen tekstsamling representert ved en rekke ulike fil- og tekstformater, og til sist at materialet er ubalansert (for mye avistekst, for lite annet).

Det er dermed usikkert hvor fruktbart det er å bruke ressurser på å avklare eller reforhandle rettigheter med hver enkelt tekstleverandør for å sikre gjenbruk og allmenn tilgjengelighet av dette tekstmaterialet, spesielt ikke i en situasjon hvor det allerede finnes store tekstmengder hos offentlige aktører som Nasjonalbiblioteket. Den overordnete konklusjon må dermed bli at tekstkorpuset har begrenset gjenbruksverdi.