Leksikalsk database for dansk

Om databasen

Den leksikalske databasen skildra under vart utvikla av firmaet Nordisk språkteknologi holding AS (NST), som gjekk konkurs i 2003. I 2006 kjøpte eit sameige samansett av Universitetet i Oslo, Universitetet i Bergen, Noregs teknisk-naturvitskaplege universitet, Språkrådet og IBM konkursbuet etter NST, for å syte for at dei språklege ressursane som NST hadde utvikla, vart ivaretekne. Nasjonalbiblioteket fekk i oppdrag frå Kulturdepartementet å byggje opp ein språkbank for norsk i 2009, og starta dette arbeidet i 2010.

Ressursane etter NST vart overførde til Nasjonalbiblioteket i mai 2011, og dei vert no gjort tilgjengelege i Språkbanken, førebels utan vidare handsaming. Språkbanken er open for attendemeldingar og forslag frå brukarane til korleis ressursane kan verte forbetra, og tek òg gjerne imot forbetra versjonar av databasen som brukarar ønskjer å dele med andre brukarar gjennom Språkbanken. Respons og attendemeldingar kan sendast til sprakbanken@nb.no.

Teksta i skildringa under er i sin heilskap skrive av Gisle Andersen, og henta frå rapporten *Gjennomgang og evaluering av språkressurser fra NSTs konkursbo*. Denne rapporten vart skriven i 2005 før sameiget (sjå over) kjøpte konkursbuet etter NST, og er ein fagleg og teknisk gjennomgang av ressursane. Teksta er tillempa situasjonen slik han er i dag. Gisle Andersen har gitt Språkbanken løyve til å nytte teksta. Den nemnde rapporten kan lastast ned i sin heilskap saman med annan informasjon via lenkja http://www.nb.no/sbfil/dok/dok.tar.gz.

Legg merkje til at den generelle skildringa av databasen i stor grad tek utgangspunkt i den norske databasen som NST utvikla parallelt med den danske.

Nasjonalbiblioteket, juni 2011

1. Generelt om NSTs leksikalske databaser

NSTs leksikalske databaser for norsk, dansk og svensk 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 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. Den danske transkripsjonen ble utført ved Center for Sprogteknologi i København.

Språkteknikernes arbeid har i hovedsak bestått i manuelt å

- transkribere ordene i henhold til gjeldende konvensjoner,
- 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 eventuelt ekspandere ordene, og
- 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 Lernout & Hauspies (L&H) proprietære transkripsjonssystem kalt L&H+. Etter bruddet med L&H ble transkripsjonene konvertert til det nøytrale fonetiske alfabetet Speech Assessment Methods Phonetic Alphabet SAMPA; 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 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.

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.

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.

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 de 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 leksikonposten 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

Felt	Feltnavn	Obl	Beskrivelse
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 i leksikonfeltene 1-2 som beskrevet i tabellen på neste side:

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 i *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 betydning 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 leksikonet. Prinsippene for leksikalsk dekomponering er for øvrig omtalt i hvert enkelt transkripsjonskonvensjonsdokument (se nedenfor).

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. NSTs danske database

3.1 Generell beskrivelse

Bruk lenken http://www.nb.no/sbfil/leksikalske_databaser/leksikon/da_leksikon.tar.gz for å laste ned NSTs danske leksikon. Følgende nøkkeltall gjelder for NSTs danske leksikon:

	Totalt	Prosentdel
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 datasettene i tabellen på neste side.

En tilhørende inspeksjonsfil *dan030224NST.pron_inspect.OUT* inneholder en mer detaljert kvantitativ fortegnelse over leksikonets innhold. Tekstfilen ligger sammen med annen dokumentasjon i arkivet http://www.nb.no/sbfil/dok/nst-dok.tar.gz.

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

3.2 Transkripsjonskonvensjoner for dansk

Transkripsjonene tar utgangspunkt i Københavnsdialekten og er basert på de samme prinsippene som beskrevet for norsk i avsnitt 4.2. med hensyn til suprasegmental markering og MOP-prinsippet. Retningslinjer for fonetisk transkripsjon av NSTs danske leksikon er beskrevet i filen *DA_SAMPA_transkonv.doc*. Merk at betydelig plass er viet håndteringen av *stød* og /r/-diftongering og vokalrealisasjoner i /r/-kontekst. Foneminventaret er beskrevet i dokumentet *PhonTable_danish_ipa_sampa_ibm_v.1.3.doc*. Begge de nevnte filene ligger sammen med annen dokumentasjon i arkivet http://www.nb.no/sbfil/dok/nst-dok.tar.gz.

3.3 Innkjøpte leksikalske ressurser

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

Navn	Innhold	Annotering
INSO	grunnformer: 75 000 bøyningsformer: 520 000	bøyningskode, ordklasse, morfologisk kode, sammen-setningsinformasjon
Institut for Navneforskning	for-, etter-, steds- og personnavn: 240 000	
KRAK forlag	gatenavn: 100 000	

3.4 Leksikalske verktøy

Det er bygget opp noen språkbehandlingsverktøy som automatiserer og forenkler arbeidet med leksikalske ressurser. Disse er for det meste skrevet i Perl. De kan lastes ned via lenken http://www.nb.no/sbfil/leksikalske_databaser/verktøy/dk.lex.tools.tar.gz. Merk at det finnes færre slike verktøy for dansk enn for norsk og svensk.

- *Verktøy for leksikoninspeksjon*
  Dette verktøyet sjekker at leksikonet inneholder all obligatorisk informasjon, kontrollerer at transkripsjonene kun inneholder valide tegn, samt lager statistikk.

- *Ordsammensettingsverktøy (Recompounder)*
  Verktøyet prosesserer gitt sammensetninger og tildeling av bitrykk, trykkforskyvinger osv. på grunnlag av inndata bestående av enkeltleddenes ortografi og transkripsjon.

- *Verktøy for grafem-til-fonem-konvertering (G2P)*
  Verktøyet konverterer ord fra ortografi til fonetisk representasjon.

- *Transkripsjonskorreksjonsprogram*
  Verktøyet gjør en grunnleggende kontroll av fonetiske transkripsjoner i leksikonet.

4 Kvalitetsvurdering

Leksikalske ressurser av denne typen 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. 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 (etter konkursen). Dermed dekker ikke ordtilfanget nyere neologismer og importord. For eksempel er et ord som *tsunami* 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.

Leksikonene er kvalitetssikret slik 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.

ELRA-dokumentet http://www.spex.nl/validationcentre/d11v21.doc (avsnitt 3.6) og manualen http://www.elra.info/services/validation_manual_lexica.pdf beskriver standarden for leksikalske databaser. 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 i all hovedsak er i samsvar med internasjonal standard for slike språkressurser.