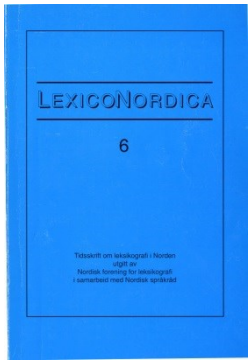
LexicoNordica

Titel: Ordboksprosjektet NORDLEXIN-N

Forfatter: Tove Bjørneset

Kilde: LexicoNordica 6, 1999, s. 35-45

URL: http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Søgbarhed
Artiklerne i de ældre LexicoNordica (1-16) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Tove Bjørneset

Ordboksprosjektet NORDLEXIN-N

A project team at the HIT centre, University of Bergen, has since 1996 been working on transferring the source material of the Swedish dictionary series LEXIN to Norwegian. The LEXIN dictionaries are customized for non-native speakers who have limited proficiency in the source language and who may also be unfamiliar with the use of dictionaries as such. The Swedish source material amounts to approx. 30.000 entries distributed on the quantitative selections MINI (approx. 5.000 entries), MIDI (approx. 8.000 entries), STORA (approx. 17.000 entries), and MAXI (approx. 30.000 entries). The objective of the project is to exploit as much of the Swedish source material as possible in order to develop a similar Norwegian database. This may in turn be linked to the already existing translations (approx. 20 languages), so that a similar dictionary series can be automatically generated in Norway. This article gives a short outline of the work carried out so far.

Bakgrunn

Den svenske Skolöverstyrelsen (SÖ) tok i 1979 initiativ til å starte et forsknings- og utviklingsprosjekt med den hensikt å produsere ordbøker som egnet seg for svenskundervisning for innvandrere. Prosjektet fikk navnet Språklexikon för invandrare (LEX-IN) og ble gjennomført i samarbeid med Statens institut för läromedel (SIL) og Statens invandrarverk (SIV). Det grunnleggende utviklingsarbeidet ble avsluttet i 1984. Deretter overtok SIV hovedansvaret for, i samråd med SÖ og SIL, å produsere ordbøker for innvandrere. Siden juli 1991 har Skolverket hatt dette ansvaret. Bakgrunnen for prosjektet var den svenske Riksdagens beslutning om at det svenske samfunnet skulle arbeide for aktiv tospråklighet og forbedre svenskopplæringen for innvandrere.

LEXIN-ordbøkene er oversiktlige og enkle å bruke, noe som gjør at også innvandrere med svært begrensede leseferdigheter eller erfaring i bruk av ordbøker kan ha nytte av dem. Ordbøkene utgis i ulike størrelser; MINI (ca. 5.000 ord), MIDI (ca. 8.000 ord), STORA (ca. 17.000 ord) og MAXI (ca. 30.000 ord). Oversettelsene omfatter per i dag rundt 20 språk, deriblant tyrkisk, albansk, arabisk, finsk, gresk, makedonsk, persisk, tigrinsk, polsk, somalisk, armensk og turabdinsk. Antall lemma i disse varierer, og flere ordbøker er under utgivelse.

Tidlig på 1990-tallet fikk samtlige nordiske land tilbud om å disponere den svenske databasen med ordboksgrunnlaget for LEXIN vederlagsfritt, for eventuell produksjon av sine nasjonale LEXIN-ordbøker. Det kongelige kirke-, utdannings- og forskningsdepartementet (KUF) var oppdragsgiver da det norske NORDLEXIN-prosjektet ble satt i gang ved HIT-senteret (Forskningsprogram for humanistisk informasjonsteknologi), avdeling Norsk termbank, Universitetet i Bergen, høsten 1996. Arbeidet hadde de første månedene status som et pilotprosjekt, der det svenske ordboksunderlaget for MINI-utvalget ble overført til norsk. I juni samme år startet arbeidet med overføringen av ordboksunderlaget for STORA-utvalget, et arbeid som pågikk til slutten av 1998. I 1999 overdrog KUF oppdragsgiverfunksjonen til Nasjonalt Læremiddelsenter (NLS), men samarbeidet med HIT-senteret fortsetter. Det norske ordboksunderlaget skal nå utvides til samme omfang som det svenske, altså rundt 30 000 ord.

De svenske ordbokspostene inneholder som et minimum:

- opplysninger om ordklasse
- fullt utskrevne bøyningsformer
- forklaring i form av omskriving og/eller synonym
- setningskonstruksjon
- uttalemarkering

– og som regel også:

- stilistisk og grammatisk kommentar
- språkeksempler og idiomer

Rundt 1 700 av de svenske lemmaene er illustrert. Illustrasjonene er inndelt i 32 emneområder (for eksempel 'familie og slekt', 'skole og utdanning' og 'natur og landskap') og er samlet i et eget vedlegg i ordbøkene. De er nå også utgitt som et eget hefte, Bildtema. Blant annet som følge av særsvensk utforming og alder kan ikke de svenske illustrasjonene brukes i Norge, men en lignende løsning vil trolig bli valgt. Foruten den primære målgruppen, som består av innvandrere og flyktninger, kan et slikt bildehefte være et godt pedagogisk hjelpemiddel også for andre elevgrupper, for eksempel døve og elever med lese- og skrivevansker.

Overføring av et ordboksgrunnlag fra svensk til norsk

Det som kanskje i størst grad særpreger det norske NORDLEXIN-prosjektet, er at det er selve kildespråksunderlaget for en ordboksserie som overføres fra svensk til norsk. Denne overføringen skjer fortløpende under gjennomgangen og vurderingen av de alfabetiserte svenske ordboksartiklene. Arbeidet kan grovt deles inn i to hovedfaser:

1. ekvivalensarbeid
2. tilrettelegging for norske forhold

Under gjennomgangen av de svenske ordboksartiklene kartlegger og etablerer vi ekvivalenser mellom de svenske ordbokspostene og de nye norske postene. I denne fasen konsentrerer vi oss om å utnytte det svenske grunnlagsmaterialet best mulig, noe som innebærer oversettelse av alle lemma, fraser, idiomer og sammensetninger. Alle ekvivalenser kontrolleres og godkjennes av en svensk filolog med lang erfaring i norskundervisning for innvandrere og flyktninger.

Hensikten med ekvivaleringsarbeidet er at vi ønsker elektronisk tilgang til de oversettelsene som allerede er gjort på svensk side. For at dette skal være mulig, må alle norske ekvivalensfelt knyttes til de svenske feltene som de er relatert til. Det satses på at en vesentlig del av oversettelsene til målspråkene skal kunne gjøres datamaskinelt, men vi må avvente oppdragsgiverens valg av målspråk før dette delprosjektet kan settes i gang.

Under tilretteleggingen for norske forhold er det naturlig å frigjøre seg noe mer fra det svenske ordboksunderlaget, for å sikre at lemmautvalget i større grad avspeiler det norske samfunnet. I denne fasen blir basen systematisk supplert med nye norske ordboksartikler. Nye fraser, sammensetninger og idiomer legges inn etter skjønn i både ekvivalerte og nye norske ordbokspostere. Til slutt legges morfologi og uttalemarkering inn i samtlige ordboksartikler.

Ordboksartiklene i LEXIN-serien inneholder flere felt som beskriver bruksområdet for lemmaet (blant annet definisjon, frase, idiom og sammensetninger). Det varierer hvilke felt som har blitt oversatt i de ulike LEXIN-ordbøkene, men vi har likevel valgt å etablere samtlige ekvivalenser vi har funnet. Det vil dessuten være av betydning for en eventuell fellesnordisk LEXIN-base at oversettelsene av det svenske underlaget er så fullstendige som mulig. Vi har tatt utgangspunkt i det svenske postformatet, men bearbeidet det noe og utviklet et eget postformat i en ORACLE-base. Ordbokspostene har, noe grovt skissert, denne strukturen:

S-lem svensk lemma
E-lem norsk lemma

S-kat	svensk ordklasse
E-kat	norsk ordklasse
S-def	svensk definisjon/synoym
E-def	norsk definisjon/synonym
S-til	svensk tilleggsdefinisjon (ved samfunnstermer)
E-til	norsk tilleggsdefinisjon (ved behov)
S-eks	svensk frase
E-eks	norsk frase
S-idi	svensk idiom med forklaring
E-idi	norsk idiom med forklaring
S-div	henvisning til svensk illustrasjon
E-div	henvisning til illustrasjon ikke etablert
S-sms	svensk sammensetning
E-sms	norsk sammensetning
S-mor	fullt utskrevne bøyningsformer
E-mor	fullt utskrevne bøyningsformer (med alle lovlige varianter)
S-utt	uttalemarkering med ordinære bokstaver, markering av tonem
E-utt	uttalemarkering med ordinære bokstaver, ikke markering av tonem

S-feltene inneholder det svenske ordboksgrunnlaget, mens E-feltene er opprettet av oss og skal fylles med ekvivalert norsk materiale. Alle ordboksposter og felt har sine unike koder med henblikk på styringen av det forstående datamaskinelle oversettelsesarbeidet. Om det ikke eksisterer noen ekvivalent for det svenske lemmaet, inaktiveres ordboksposten, slik at den ikke lenger er en del av den norske databasen. Mangler det en norsk ekvivalent for et av de andre feltene, inaktiveres feltet ved at det settes inn en bindestrek. Eventuelt legger vi inn et særnorsk felt med innhold som ikke skal knyttes til noe svensk felt, og som da må oversettes på nytt til det aktuelle målspråket. Den ekvivalerte ordboksartikkelen under kan illustrere dette. Ordboksartikkelen er fullstendig ekvivalent. Det svenske og det norske begrepet er det samme, men definisjonen er utdypet:

S-lem	**glas**
E-lem	**glass**
S-kat	subst.
E-kat	subst
S-def	ett hårt och genomskinligt ämne
E-def	–
N-def	et hardt, knusbart og gjennomsiktig materiale
S-div	bild 7:14
E-div	
S-sms	glas~tillverkning -en
E-sms	glass~produksjon -en
S-mor	glaset glas glasen
E-mor	glasset glass glassene (el glassa)

N-alt	glas
S-utt	gla:s
E-utt	glas

Ekvivalens

NORDLEXIN-N-prosjektet er ikke et oversettelsesprosjekt i vanlig forstand, men mer et *ekvivaleringsprosjekt*. Å kartlegge og etablere ekvivalenser for lemma-feltene har vært det mest sentrale i dette arbeidet. I *Nordisk leksikografisk ordbok* (1997) defineres 'ekvivalens' slik:

> *semantisk og funksjonsmessig overensstemmelse mellom ord eller uttrykk i to eller flere språk.*

Begrepene kan ofte være vagere og mer upresise i allmennspråk enn i fagspråk. Den typiske termgruppen består hovedsakelig av substantiv, mens dette ordboksprosjektet i større grad har hatt leksem fra alle ordklasser. *Fullekvivalens*, det vil si fullstendig samsvar i tegninnhold og betydningsomfang, forekommer sjelden i allmennspråk, men oftere i fagspråk. Det er kartlagt flere fullekvivalenser enn forventet i prosjektet. Ved partiell ekvivalens er det delvis likhet mellom kildeord og målord. Forskjellen kan blant annet vise seg ved at begrepet i målspråket har færre karakteristiske trekk enn begrepet i kildespråket. Eksempelvis kan det svenske lemmaet 'betyg' i mange tilfeller oversettes til 'karakter', men vi må tenke generelt og begrepsorientert. I tillegg til denne betydningen, dekker 'betyg' også det norske ordet 'vitnemål'. Vi kan derfor ikke ekvivalere det svenske lemmaet med 'karakter'. Fordi det ikke finnes noen norsk overterm som dekker både 'karakter' og 'vitnemål', kan vi heller ikke sette inn en norsk ekvivalent for 'betyg'. Vi må i stedet inaktivere den svenske posten og opprette to nye ordboksposter for 'karakter' og 'vitnemål'. Det er også kartlagt en viss mengde av nullekvivalenser. Ikke sjelden mangler norsk fullstendige eller partielle ekvivalenter til de svenske leksikalske størrelsene. Som regel inntreffer dette i møte med kulturspesifikke svenske begreper, som for eksempel med 'De Aderton' (medlemmene i Svenska Akademien).

Nye, norske lemma i NORDLEXIN-N

Svensk og norsk språk er mer ulike enn man gjerne vil tro. Tilsynelatende identiske ord viser seg gjerne å være falske venner (jf. det svenske anglolånet 'freestyle', som på norsk betyr 'walkman') eller tilsløre ofte viktige betydningsnyanser og -forskjeller. Ord som er synonyme i ett av språkene kan ha klart atskilte betydninger i det andre språket. I en del tilfeller har slike og lignende ulikheter gjort det nødvendig å lage nye norske poster. Vi har arbeidet i nær dialog med fagfolk i ekserperingsdelen av dette arbeidet.

Samfunnsinstitusjoner med tilhørende stillingskategorier og funksjoner er ofte ulikt organisert i Sverige og Norge. En rekke ord som er knyttet til det svenske samfunnet og samfunnssystemet har vi valgt å inaktivere i den norske versjonen av ordboksgrunnlaget. I stedet er tilsvarende norske institusjoner, sosiale ordninger og lignende lagt inn som nye oppslag. Videre har det vært nødvendig å legge inn en del ord som refererer til nyere og mer moderne fenomener. Dette gjelder blant annet dataterminologi og uttrykk knyttet til samfunns- og politiske forhold fra de siste par tiårene. Det er også ekserpert ord og uttrykk fra brosjyrer og foldere som er utgitt av kommunale etater (særlig helse- og sosialvesenet), apotek, bank, postkontor, arbeidskontor og offentlige kommunikasjonsbedrifter, for å fange opp flest mulig ord som man møter i hverdagslivet. Tilsvarende er det lagt inn ord og uttrykk knyttet til skolefag, skoleverk og lignende, og dessuten en del ord og uttrykk knyttet til sport og fritidsaktiviteter. Vi har valgt å utelate opplagt tidsavgrenset informasjon, med unntak av blant annet de største norske politiske partiene og organisasjonene.

Prosjektnormal

Alle former innenfør læreboknormalen skal være representert i ordboksunderlaget for NORDLEXIN-N. De praktiske konsekvensene av dette er blant annet et nærmest uendelig mangfold av valgmuligheter: Hvilken rekkefølge bør de ulike formene presenteres i? Hvilke former bør vi bruke i definisjoner og eksempler? Etter hvert kom vi fram til at det ikke ville være riktig å bruke bare én variant av norsk bokmål. I ordboksprosjekter som dette er det etter vår oppfatning viktig å ikke tilsløre de relativt store variasjonsmulighetene, men tvert imot å gjøre dem synlige for brukeren.

Vi valgte likevel å etablere og operere med en allmennspråklig prosjektnormal til bruk i lemma- og definisjonsfeltene. Prosjektnormalen kan karakteriseres som et relativt moderat bokmål. I andre felt har vi bevisst vekslet mellom de ulike tillatte formene i læreboknor-

malen. Innbyrdes inkonsekvens har ikke blitt vurdert som problematisk, men som ellers i bokmål er det gjerne konkrete substantiver og visse stilistiske kontekster som i særlig grad motiverer valg av hokjønnsord og andre "radikale" former. Til sjuende og sist blir dette likevel bare et spørsmål om skjønn:

monoftong - diftong
løs (laus)
sen (sein)
bløt (blaut)

hv/kv
hvit (kvit)
kvass (hvass)

y/ju
syk (sjuk)
tykk (tjukk)
e/jø
selv (sjøl)
mel (mjøl)
bjørk (bjerk)

u/o
hule (hole)
hugg (hogg)

hard versus myk konsonant
ligne (likne)

andre
fram (frem)
abbor (åbor)
albue (alboge)
aske (oske)
bånd (band)
vei (veg)
hammer (hammar)

verbbøying, partisipp
kastet (kasta)
stripete (stripet)

substantivbøying
bevere (bevrer) - beverne (bevrene)
slottene (slotta)
kontorer (kontor)

felleskjønn eller hunkjønn
antennen (antenna)
adressen (adressa)

sammenskriving eller særskriving
allting (all ting)
riktignok (riktig nok)
visstnok (visst nok)

kort eller lang form
bestride (bestri)
be (bede)
blø (bløde)
rå (råde)

tegnsetting i forkortelser
ca (ca.)
f eks (f.eks.)
m oh (m o.h.)

store eller små bokstaver i leksikaliserte forkortelser
lp (LP)
ekg (EKG)
pc (PC)
aids (AIDS)

norsk eller engelsk skrivemåte
teip (tape)
vaier (wire)

Valgfrie former

De ulike formene er atskilt med mellomrom. Sidestilte former settes i vanlige parenteser og innledes med "el" for "eller". For lemmaer med flere sidestilte former har vi hittil valgt å sette de alternative formene i

parentes. Den store valgfriheten i norsk gjør imidlertid at denne løsningen fungerer vesentlig dårligere hos oss enn i Sverige, og vi ønsker derfor å endre dette oppsettet.

I morfologifeltet er de valgfrie formene lagt inn i samme rekkefølge som de ulike sidestilte formene av lemmaet. 'Beruset', 'berusa' og 'berust' er sidestilte former. 'Beruset' er valgt som hovedinngang, og 'berust¨' og 'berusa' er lagt inn i feltet for alternative former. I morfologifeltet settes de finitte formene opp i samme rekkefølge:

S-lem	**berusad**
/.../	
S-eks	en kraftigt berusad yngling
E-eks	en svært berusa ungdom
N-eks	hun var beruset av suksessen
S-mor	berusat berusade
E-mor	beruset (el berust berusa) berusede (el berusete beruste berusa)
N-alt	berusa berust
S-utt	berU:sad
E-utt	beru55:set

I tilfeller der lemmaet bare har én lovlig variant innenfor læreboknormalen, men det er flere valgfrie bøyningsformer, har vi som et grunnprinsipp valgt å sette de mest konservative formene først:

S-lem	**abonnemang**
E-lem	**abonnement**
/.../	
N-eks	jeg avsluttet abonnementet på lokalavisen
S-sms	årsabonnemang
E-sms	årsabonnement
S-sms	abonnemangs~avgift -en
E-sms	abonnements~avgift -en/-a
S-mor	abonnemanget abonnemang abonnemangen
E-mor	abonnementet abonnementer (el abonnement) abonnementene (el abonnementa)
S-utt	abånemAN:
E-utt	abonema5ng

Vi har forsøkt å rasjonalisere med plassen i de norske morfologifeltene, der dette har vært mulig. Når variasjonen mellom ulike former ikke gjelder bøyningsparadigme, men annen variasjon, for eksempel monoftong/diftong, mener vi at det er tilstrekkelig å illustrere bøyningen med den varianten som er oppført i lemmafeltet. Ordene *sen*

og *sein* har identisk bøyning, og det er derfor ikke nødvendig å gjenta informasjonen om at også diftong er tillatt, i morfologifeltet.

S-lem **sen 2**
E-lem **sen**
/.../
S-mor sent sena
E-mor sent sene
N-alt sein
S-utt se:n
E-utt se:n

Uttale

Det finnes ingen normert uttale for norsk, og heller ikke finnes det noe 'høyspråk'. Det har etter hvert blitt akseptert å snakke sin egen dialekt i alle sammenhenger, uansett posisjon og tilholdssted. Det nærmeste vi kommer en standardisert uttale, er kanskje scenespråket som brukes på teater og lignende steder, men dette høres ofte stivt og unaturlig ut for de fleste. I vår statlige nyhetskanal NRK må journalistene snakke bokmål eller nynorsk i nyhetssendingene, men uttalen kan likevel være svært forskjellig, avhengig av hvor i landet de kommer fra. På bakgrunn av dette er det etter vår oppfatning ikke mulig å gi en nøyaktig uttalemarkering for moderne norsk. En uttalemarkering som skal være funksjonell for flest mulig i målgruppen, må nødvendigvis være så generell at den ikke blir direkte misvisende for svært mange av de norske dialektene. Likevel må den gi mer informasjon enn selve skriftbildet gir.

Der det har vært nødvendig å velge mellom to uttalemåter med geografisk variasjon, har vi valgt den som dominerer i Østlandsområdet, fordi de største delene av målgruppen bor der. Er to uttalemåter utbredt over hele landet, er begge tatt med.

Det er ikke brukt IPA-tegn i uttalemarkeringene, da disse tegnene ble vurdert som for avanserte for den primære målgruppen. I stedet brukes den vanligste bokstaven eller bokstavkombinasjonen for en norsk lyd. Hver lyd er vanligvis representert av én bokstav i uttalemarkeringen. Diftonger markeres med bue.

Kvantitet og aksent

Kvantitet er bare markert for vokalene. Markeringen består av et kolon direkte etter den lange vokalen, fordi lengden på konsonantene kan utledes fra lengden på vokalene:

lete etter noe	[le:te]
flyet skal snart lette	[lete]

Trykk markeres med et punkt under den aksentuerte vokalen, med unntak av i enstavelsesord. Hovedtrykk markeres i alle posisjoner. Bitrykk markeres ikke:

boksta5:v
arbæider

Tonem

Norsk har to tonem (også kalt 'tonelag'). Det er ulik uttale på [1]tanken (bestemt form av substantivet 'tank') og [2]tanken (bestemt form av substantivet 'tanke'). Dette markeres likevel ikke, fordi det er store dialektvariasjoner med hensyn til hvilket tonem et ord har, og også med hensyn til hvilke ord som har tonemvariasjoner eller ikke.

Ordklasseinndeling

Vi har tatt utgangspunkt i *Norsk referansegrammatikk* og bruker disse ordklassene:

Substantiv	(en gammel **mann**)
Pronomen	(**han** spurte **henne** om de skulle gifte **seg**)
Determinativ	(**en gammel** mann)
Adjektiv	(en gammel mann)
Adverb	(det regner **ikke nå**)
Verb	(jeg **heter** Nina og **bor** i Norge)
Preposisjon	(jeg bor **i** Bergen)
Interjeksjon	(**huff!**, **æsj!**, **au!**)
Konjunksjon	(han var snill **og** grei)
Subjunksjon	(han spurte **om** jeg kom)

Sluttkommentar

Vi har ofte blitt spurt om årsaken til at vi går via et svensk ordboks-underlag i arbeidet med å lage norske ordbøker for minoritetsspråklige innvandrere. Som kjent kan ordboksproduksjon foregå på ulike måter, og NORDLEXIN-N-prosjektet kan kanskje sies å være tuftet på en kombinasjon av disse. Ved å overføre det svenske ordboksunderlaget for LEXIN-ordbøkene til norsk, satser vi på å oppnå en tidsgevinst med henblikk på elektronisk tilgang til de rundt 20 målspråkene som LEXIN til nå er oversatt til. Videre bør det ligge et ikke ubetydelig kvalitetssikringsaspekt i det å bygge på et svensk ordbokskonsept som har fungert godt i en svært differensiert målgruppe gjennom snart 20 år. På dette grunnlaget håper vi å kunne utvikle funksjonelle og brukervennlige ordbøker i både trykt og elektronisk form for inn-vandrere også i Norge.

Litteraturliste

Beijer, Maj og Kiros Fre Woldu 1997: *Detta är LEXIN. Lexicon för invandrare*. Stockholm: Skolverket.

Bergenholtz, Henning m.fl. 1997: *Nordisk leksikografisk ordbok.* Oslo: Universitetsforlaget.

Faarlund, Jan Terje m.fl. 1997: *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Svensén, Bo 1987: *Handbok i lexikografi. Principer och metoder i ordboksarbete*. Stockholm: Norsteds.