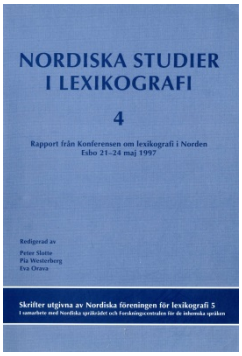
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Hybris - nemesis - balance. Problemer med genbrug af ordbogsdata set fra Den Danske Ordbog

Forfatter: Ebba Hjorth

Kilde: Nordiska Studier i Lexikografi 4, 1997, s. 189-193
Rapport från Konferens om lexikografi i Norden, Esbo 21.-24. maj 1997

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



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) 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.

Ebba Hjorth

Hybris – nemesis – balance. Problemer med genbrug af ordbogsdata set fra Den Danske Ordbog

According to the original plans, The Danish Dictionary (DDO) was to build upon already existing machine-readable dictionary data as well as make use of quotations from a machine-readable corpus, which had been compiled by The Danish Dictionary and consists of 40 million running words. However, when the machine-readable dictionary data was compared with the language usage in the large text corpus, the original plans had to be revised and the intention of re-using existing dictionary data was abandoned. The choice of words as well as the orthographic, morphological, semantic and syntactic descriptions had to be revised radically in order to make the language coverage more adequate. The dictionary had hereby changed status from being a **corpus-backed** dictionary to being a **corpus-based** dictionary.

Hybris

Da det danske kulturministerium og Carlsbergfondet i 1991 tilsammen bevilgede 30 millioner danske kroner til fremstilling af Den Danske Ordbog (DDO) i seks bogtrykte bind, var det på grundlag af en nøje udarbejdet plan for arbejdet. Planen skulle sikre at budgettet ikke blev overskredet, og at ordbogen blev færdig i løbet af forholdsvis kort tid (7 år), således at afstanden mellem ordbogens eksempelmateriale der hentedes fra et 40 millioner ord stort tekstkorpus, og ordbogens publiceringstidspunkt ikke blev for stor. Planen byggede på en høj grad af computerisering og på en udstrakt grad af genbrug af allerede eksisterende maskinlæsbare ordbogsdata. Grundideen var altså – lidt firkantet udtrykt og sådan som vi også tidligere har beskrevet det ved konferencer i leksikografiske sammenhænge – at et råmanuskript fremstillet af eksisterende ordbogsdata skulle eksemplificeres med citatater fra det tekstkorpus på 40 millioner ord som ordbogen etablerede. Ordbogen skulle altså være **en korpusstøttet ordbog**. Ordbogens tidsplaner og budget blev opstillet ud fra disse ideer.

Ordbogen blev både over for de bevilgende myndigheder, over for pressen og over for fagkolleger lanceret med udtryk som "Verdens hurtigste ordbog", "ordbog til tiden" og andre lignende overmodige udsagn.

"ny dansk ordbog på otte år"
(Jyllandsposten 29. juni 1991)

"Den Danske Ordbog. Verdens hurtigste.
Det bliver noget nær verdensrekord"
(Aalborg Stiftstidende 5. juli 1991)

Nemesis

1. september 1991 gik arbejdet på DDO i gang. Opbygningen af korpusset fyldte den første tid. De 40 millioner ord blev samlet i hus. Nogle af ordene var maskinlæsbare fra begyndelsen, andre blev gjort det enten ved skanning eller ved indtastning. Alle de 45.000 tekstprøver blev forsynet med kildeoplysninger, oplysninger om forfatteren/sprogbrugeren, kommunikationssituation, genre etc.

En række forsøg med genbrugsredigering blev derefter foretaget. På næsten alle områder i ordbogen havde vi planer om genbrug, næsten alle oplysningstyperne skulle hente data fra andre ordbøger. I det følgende gives en oversigt over de vigtigste oplysningstyper der skulle hente informationer fra andre ordbøger:

Oplysninger om **ortografi** skulle hentes fra Retskrivningsordbogen (RO). Oplysninger om **bøjning** skulle ligeledes hentes fra RO. **Udtalen** skulle findes i én af de to eksisterende udtaleordbøger. **Udvælgelsen af ordforrådet** skulle ske på baggrund af to store fremmedsprogsordbøger plus RO. De to fremmedsprogsordbøger er Vinterberg og Bodelsen: Dansk-engelsk Ordbog (V&B) og Blinkenberg og Høybye: Dansk-fransk Ordbog (B&H). Begge disse store fremmedsprogsordbøger var netop udkommet i nye udgaver, og vi havde adgang til maskinlæsbare versioner af de to værker. **Den semantiske beskrivelse** skulle først og fremmest støtte sig til de to store fremmedsprogsordbøger og til Ordbog over det danske Sprog (ODS) og dets Supplement. **Konstruktionsoplysninger** skulle findes i Erik Bruun: Dansk Sprogbrug.

Desværre viste det sig meget snart at det at man tog udgangspunkt i et råmanuskript fremstillet på baggrund af de eksisterende ordbøger, stillede sig hindrende i vejen for beskrivelsen af det sprog der kunne iagttages i korpus. Korpus viste sig nemlig at være et fremragende empirisk grundlag for en adækvat beskrivelse af moderne dansk. Det ordbogsprodukt der ville blive resultatet af genbrugsstrategien, ville blive så meget anderledes end det produkt der kunne fremstilles på baggrund af det nyopbyggede korpus, at vi ikke længere fandt det forsvarligt at fortsætte arbejdet efter den oprindelige plan.

Lad mig give nogle eksempler:

Ordudvælgelse:

En ordudvælgelse på baggrund af V&B, B&H og RO ville give problemer med både for mange ord og for få ord:

En lang række ord som de fleste danskere kender og af og til også bruger, optræder ikke i hverken ODS, i Nudansk Ordbog, i Vinterberg og Bodelsens engelsk-danske ordbog eller i Blinkenberg og Høybyes store dansk-franske ordbog. Disse ord er vi blevet opmærksomme på gennem søgninger og optællinger i vores korpus. Det drejer sig fx om ord som: *amatøragtig, babytøj, brudevals, bykort, dersens, havegang, indvirken, kødsovs, langlemmet, litervis, mavemuskel, møjunge, nettohusleje* og mange flere. Når disse ord ikke er med i V&B og B&H, skyldes det ikke at disse ordbøger er dårlige ordbøger, men snarere deres tilblivelseshistorie. De er nemlig – selv om de er kommet i nye udgaver – afhængige af ODS, hvor ordudvælgelsen er manuel og derfor udsat for tilfældigheder. Oversættelsesordbøgerne repræsenterer altså via sin afhængighed af ODS et væsentligt ældre sprogtrin end deres udgivelsesår tilsiger.

I oversættelsesordbøgerne og i RO er der en del såkaldte "ordlig". Det er ord som står i ODS, men som ikke kan siges at tilhøre det moderne ordforråd, altså ord som mange danskere kender

og også af og til selv bruger. En søgning og optælling i DDOs korpus får os til at tage en dyb indånding og smide disse lig ud. Et eksempel er ordet *huemoder* der betyder 'den person der bærer et dåbsbarns hue under selve dåbshandlingen' eller *gramsepose* der betyder 'pose hvoraf man paa lykke og fromme trækker et lod ell. en gevinst' (ODS).

At basere sit ordudvalg på eksisterende ordbøger ville altså være at beskrive et ældre sprog end det der kommer til syne i vores korpus. En række opslag i DDO ville give et negativt resultat med irritation over værket til følge, og en del plads ville blive optaget af ord der ikke tilhører ordforrådet i den periode som vi beskriver.

Semantik:

Også den semantiske distribution, den semantiske beskrivelse og de enkelte betydningers frekvens kommer i vanskeligheder hvis man bygger på eksisterende ordbøger. Sammenlignet med den sproglige virkelighed som korpus repræsenterer, har de tilgængelige ordbøger dels for mange, dvs. forældede betydninger, dels for få, dvs. manglende nye betydninger. Et eksempel:

Ordet *genmæle* har i følge vores korpus en betydning svarende til 'berigtigelse'. Denne betydning er ikke registreret i eksisterende ordbøger, men er så frekvent i vores korpus at vi ikke kan undlade at beskrive betydningen. Bygger vi udelukkende på genbrug, kan vi altså ikke være sikre på at opdage denne nye betydning, for registreringen af den kræver en grundig analyse af korpusforekomsterne, især hvis redaktøren af det pågældende ord ikke selv kender denne nye betydning (spørgsmålet om norm over for usus skal ikke berøres i denne sammenhæng).

På tilsvarende måde viser korpusanalyserne fravær af betydninger som er beskrevet i de eksisterende ordbøger.

Dertil kommer at ordningen af betydningerne i de eksisterende ordbøger svarer til ODS's historiske rækkefølge. Det giver vanskeligheder for DDO der som hovedregel ordner betydningerne efter faldende frekvens.

Altså igen: den sproglige virkelighed i korpus svarer ikke til de eksisterende ordbøgers beskrevne virkelighed.

Konstruktionsoplysninger:

Vi havde oprindelig tænkt os at hente konstruktionsoplysninger fra Erik Bruun: Dansk Sprogbrug. Også hér kommer der uoverensstemmelse mellem korpus og ordbogens oplysninger. På én gang rummer konstruktionsordbogen for mange og for få oplysninger. Konstruktioner der beskrives i ordbogen, er ikke længere gængse, findes i hvert fald ikke i vores korpus og kan derfor ikke beskrives i DDO som gængs moderne dansk, ligesom korpus viser måder at konstruere sætninger eller vendinger på som ikke er beskrevet i Dansk Sprogbrug. At beskrive de mest almindeligt forekommende valens- og konstruktionsmønstre i moderne dansk ordnet efter frekvens kan kun gøres på forsvarlig måde efter en analyse af korpus.

Ortografi, bøjning:

Også på disse to områder har korpus korrektioner til den eksisterende retskrivningordbog. DDO er normativ og beskrivende på én gang og kan derfor ikke "nøjes" med at videregive den norm som

Retskrivningsordbogen er udtryk for. Både med hensyn til bøjning og stavning viser DDO's korpus en side af moderne dansk som ikke er beskrevet i RO. Spørgsmålet om deskriptiv over for normativ leksikografi lades igen ude af diskussionen.

Udtale:

Med hensyn til udtale indeholder DDO's korpus i følge sagens natur ikke korrektioner til de eksisterende ordbogsværker. Så på dette område bliver graden af genbrug forholdsvis stor. En datafil med udtaleangivelserne fra Gyldendals Dansk Udtale er blevet konverteret fra IPA- til Dania-lydskriften. Disse udtaleangivelser korrigeres af en redaktør på DDO.

Nået til disse erkendelser, måtte DDO skifte status fra korpusstøttet ordbog til korpusbaseret ordbog. Vi måtte arbejde udelukkende på grundlag af korpus når det gælder de semantiske oplysninger og konstruktionsoplysninger, mens vi med hensyn til ortografi, bøjning, udtale og ordudvælgelse i høj grad på baggrund af korpus måtte korrigere og supplere eksisterende ordbogsværkers oplysninger.

Balance?

Disse erkendelser kunne naturligvis få dramatiske konsekvenser for ordbogens tidsforbrug. At gennemanalysere et korpus på 40 millioner ord kræver nye metoder og oplæring af redaktørerne. For at kunne undersøge om de tildelte bevillinger kunne række til en færdiggørelse af ordbogsmanuskriptet til tiden, omfordelte vi resurserne fra eksterne fagkonsulenter til interne faste redaktører. Desuden udviklede vi et administrationssystem som kunne følge redaktionsprocessen meget nøje. Da administrationssystemet var delvis færdigudviklet og redaktionsreglerne var på plads og redaktørerne ensrettet (så meget som man nu kan ensrette redaktører), satte vi tidsstudier i gang. Og det viste sig at det ikke var muligt at indhente den tid der var gået med ændringsprocessen og udviklingen af nye metoder og arbejdsprocedurer. Vi måtte altså gå den tunge gang til bevillingsgiverne og forklare dem situationen. Det er lykkedes for os at vinde bevillingsgivernes accept af den ændrede arbejdsmetode. Både Kulturministeriet og Carlsbergfondet har ønsket at få den korpusbaserede ordbog, og DDO har fået forøget sine samlede bevillinger med 40%. Pengene bruges ikke til at ansætte flere redaktører, men til at forlænge redaktionsperioden. Udsendelsen af DDO udsættes til 2002. Vi er i øjeblikket 10 redaktører, og vi mener det er en passende størrelse på en redaktion. Hver gang vi har haft ledige stillinger (hvilket vi heldigvis ikke har haft ret tit), har vi haft kvalificerede ansøgere, men (på en enkelt undtagelse nær) ikke ansøgere med leksikografisk erfaring, og jeg behøver næppe i denne forsamling at sige at det tager tid at oplære en leksikograf.

DSL som udgiver DDO, har altså været i en situation hvor det er lykkedes at skaffe flere midler til et projekt som vi selv havde sat rammerne for. Det var næppe sket på et privat forlag, men det viser at de enkelte sprogsamfund har brug for en offentligt støttet leksikografisk institution med faste bevillinger, der kan udføre den leksikografiske grundforskning, som andre ordbogsprojekter kan hente oplysninger fra. For selv om DDO nu har fået øget sine bevillinger, har vi stadigvæk travlt. De nye tidsplaner levner ikke plads til mange svinkeærinder. Hver dag skal der redigeres og gennemlæses nøje fastsatte mængder af ordbogslinjer – og vi tør ikke én gang til komme til bevillingsgiverne og bede om flere penge – for vi har endnu engang lovet "ordbog til tiden".

Litteratur

B&H = Andreas Blinkenberg og Poul Høybye: Dansk-fransk ordbog. 4. udgave. København 1991.

Bruun, Erik: Dansk Sprogbrug. En stil- og konstruktionsordbog. 2. udgave. Gyldendal. København 1995.

Den Store Danske Udtaleordbog af Lars Brink, Jørn Lund, Steffen Heger og J. Normann Jørgensen. Munksgård. København 1991.

Hansen, Peter Molbæk: Dansk Udtale. Udtaleordbog. Gyldendal. København 1990.

Nudansk Ordbog = Politikens Store Nye Nudansk Ordbog. 1. udgave. København 1997.

ODS = Ordbog over det danske Sprog 1–28. Udgivet af Det Danske Sprog- og Litteraturselskab. København 1919–1956.

RO = Retskrivningsordbogen. 2. udgave. Dansk Sprognævn. København 1996.

ODS-S = Supplement til Ordbog over det danske sprog 1–2. Udgivet af Det Danske Sprog- og Litteraturselskab. København 1992–1994.

V&B = Hermann Vinterberg og C.A. Bodelsen: Dansk-engelsk Ordbog. 3. udgave ved Viggo Hjørnager-Pedersen. Gyldendal. København 1990.