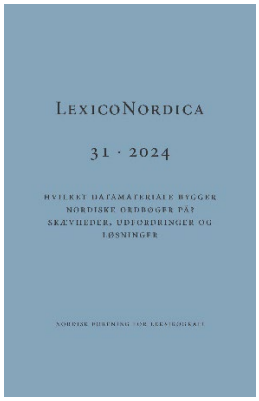
LexicoNordica

Titel:	Fra ‘sandheden om sproget’ til et opgør med stereotyper: ikke-korpusbaserede metoder til leksikografiske beskrivelser af kontroversielle ord
Forfatter:	Sanni Nimb
Kilde:	LexicoNordica 31, 2024, s. 151-174
URL:	https://tidsskrift.dk/lexn/issue/archive



Fra 'sandheden om sproget' til et opgør med stereotyper: ikke-korpusbaserede metoder til leksikografiske beskrivelser af kontroversielle ord

Sanni Nimb

In the last decade, the lexicographic community has begun to realize that the description of sensitive lemmas concerning, e.g., gender, sexuality, and ethnicity cannot be based on the usual corpus linguistic methods due to the risk of biases and stereotypes in corpora. To be able to assess the extent to which sensitive lemmas are derogatory, offensive, or politically incorrect, we instead suggest a method where the lemmas are annotated by a group of language users representing different genders and ages. Exactly which lemmas are considered sensitive by (sometimes only parts of) the language community is also difficult to determine. This article presents an approach where information from existing usage labels in a dictionary is combined with thesaurus information.

1. Indledning

Da korpusleksikografiske metoder blev taget i anvendelse for mere end 30 år siden, blev de i de første mange år anset som meget velegnede til at undgå subjektive og holdningsprægede ordbogsoplysninger. I de senere år er der opstået en større bevidsthed om at dette ikke altid er tilfældet. Denne artikel diskuterer udfordringerne ved at basere beskrivelser af kontroversielle ord, det vil sige ord relateret til sensitive emner som fx køn, seksualitet, etnicitet og handicap, på korpuslingvistiske metoder – udfordringer der skyldes omfanget af stereotyper og bias i tekstkorpora, næsten uanset hvor balanceret de er sammensat. Artiklen fremlægger først en metode til at indkredse kontroversielle ord, idet også ord der kun

har potentiale til at blive kontroversielle, inkluderes. Den giver efterfølgende et bud på hvordan man kan sikre at beskrivelserne af de udvalgte ord ikke alene baseres på leksikografens egen intuition når stereotype data fra korpus forkastes. Den nødvendige indsigt i ordenes nedsættende, krænkende eller politisk ukorrekte konnotationer kan opnås ved at lade en gruppe lingvister opmærke ordene systematisk. Det er en vigtig pointe at de pågældende lingvister repræsenterer forskellige aldersgrupper og køn, fordi der ofte er stor variation i holdninger til ord inden for de pågældende semantiske områder. De opmærkede data kan både danne udgangspunkt for et særligt 'tabu'-leksikon til brug i NLP (natural language processing) og til angivelse af sprogbrugsoplysninger i en ordbog.

Indledningsvis (afsnit 2) beskriver jeg hvordan fremkomsten af korpusleksikografiske metoder blev modtaget i ordbogsmiljøer i 1980'erne og 90'erne, og hvordan det bl.a. influerede på redigeringen af *Den Danske Ordbog* (DDO; 2003-05) i 1990'erne og 00'erne. I afsnit 3 diskuterer jeg de i dag anerkendte problemer ved at bruge korpuslingvistiske metoder til beskrivelsen af sensitive ord. I afsnit 4 diskuterer jeg mulige strategier før jeg i afsnit 5 og 6 fremlægger metoder til at indkredse, henholdsvis annotere det sensitive ordforråd.

2. Korpusleksikografiens fremkomst og 'sandheden om sproget'

I ældre danske ordbøger, hvor beskrivelserne var baseret på få indsamlede belæg og leksikografens egen intuition, skal man ikke lede længe før man finder eksempler på holdningsprægede beskrivelser af fx køn og seksualitet. I *Moths Ordbog*, der beskriver dansk i årene omkring 1700, skinner et negativt syn på kvinder i samtidens mandsdominerede samfund fx tydeligt igennem. Man finder således en stor overvægt af negative kollokationer og udtryk

ved lemmaet *Kvinde* ("kvinde er Satans redskab", "Ond kvinde er verre end pest, end skærsîld", "Mand mâe ei troe kvinder, ikke den allerbeste") og også ved lemmaet *Kvind-folk* ("kvindfolk er altid galne", "Kvindfolk er ei at troe" og "alt ont kommer fra kvindfolk"). Også i *Ordbog over det danske Sprog* (dansk sprog i perioden 1700-1950), der blev redigeret 200 år senere i perioden 1918-56, skinner negative holdninger vedrørende køn og seksualitet af og til igennem. En overført negativ betydning af lemmaet *Højttaler*, 'højrøstet, meget talende kvinde', nævner fx *hustru* i beskrivelsen: "(jarg., spøg.) om (meget ell. højrøstet talende) kvinde, især hustru", og en betydning af lemmaet *kæresteri* ("m. h. t. personer af samme køn") dokumenteres med følgende citat: "en vis Art af unaturligt Kæresteri mellem Kvinder, hvor det just ikke kommer til perverse Aktus, men hvor der dog kan erkendes en usmagelig Intimitet i Forholdet".

Med korpusleksikografiens fremkomst i løbet af 1980'erne og 90'erne fik man mulighed for at undgå subjektive og holdningsprægede beskrivelser. Williams (2003) beskriver det ligefrem som et revolutionerende metodeskift inden for leksikografien:

> A revolution in dictionary making came with the development of corpus linguistics, built on the contextualist view of meaning, and its transfer to lexicographical practice through the COBUILD dictionaries.

Man opfattede resultater af korpusundersøgelser som objektive data og sandheden om sproget, uanset en samtidig erkendelse af at de statistiske resultater ofte afspejlede negative og stereotype holdninger til fx køn og minoriteter. DDO, der blev redigeret i årene 1992-2004, var en af de første fuldt ud korpusbaserede ordbøger, og man redigerede ordbogen ud fra netop denne holdning. Scheuer (1995; forfatteren var it-redaktør ved projektet) analyserer fx bigrammer fra korpuslingvistiske undersøgelser i *Den Danske Ord-*

bogs korpus (40 millioner ord, bestående af tekster fra perioden 1983-1992) og konkluderer at de ofte er meget stereotype. Han konstaterer at "[d]er gives dette billede af kønnene: manden er et åndsvæsen […] Kvinden er en kødvare. Og det billede skal reproduceres i en deskriptiv ordbog", og at "[d]en deskriptive ordbog er et vigtigt historisk dokument, og DDO portrætterer samfundet og samfundsideologien som den ser ud i tiåret 1983-92" (Scheuer 1995:255). Eksempler på stereotype bigrammer som påpeges af Scheuer, er *hans kolleger*, *hans kontor*, *hans ideer*, *hans firma* over for *hendes hånd*, *hendes hår*, *hendes mor*, *hendes latter*. De stereotype bigrammer udgjorde materiale til kollokationerne i den første, trykte udgave af DDO, kollokationer der stadig findes i nutidens online-udgave af ordbogen. Man finder fx under opslagsordet *kvinde* disse kollokationer: *gifte kvinder*, *enlig kvinde*, *en smuk kvinde*, *kvinder og børn*, og under opslagsordet *mand* i stedet disse: *klog mand*, *rig mand*, *stærke mænd*, *rigtige mænd*. Fjeld (2015) påpegede mange stereotype citater i DDO, jf. fx citatet "[kvinderne] er eftergivende, indordner sig og accepterer at arbejde og leve på andres præmisser" under opslagsordet *eftergivende*.

De stereotype bigrammer og citater skyldtes ikke at DDO-korpusset var specielt ensidigt opbygget. Tværtimod lagde man i DDO-projektet meget stor vægt på at gøre det så velafbalanceret som muligt. Teksterne omfattede både aviser, ugeblade og magasiner, litteratur, private tekster i form af dagbøger og endda også talesprog. Scheuer (1995:246) skriver om indsamlingen af tekster til korpusset: "Der blev altså sat visse minimumskrav til spredningen – i praksis blev spredningen enorm". Der var også en vis balance i skribenternes køn, idet 1/3 var kvinder, og bigrammer målt på kun tekster af kvinder var i øvrigt også i lige så høj grad stereotype som dem beregnet på det samlede korpus (egen undersøgelse i 1996 da jeg var ansat som redaktør ved DDO; ikke publiceret). Dog var en væsentlig andel af teksterne avistekster. Da avistekster netop er den genre der indeholder flest stereotyper iføl-

ge Müller-Spitzer & Rüdiger (2022), var dette måske med til at øge antallet af stereotype bigrammer og citater.

3. Korpusleksikografiens særlige udfordringer: kontroversielle ord

I det seneste årti er der kommet et andet syn på reproduktion af de stereotyper der udledes statistisk af store tekstsamlinger, end det der lå bag redigeringen af DDO i 1990'erne. Inden for fagområdet leksikografi bliver emnet diskuteret bredt, fx i Fjeld (2015), hvor stereotyper i nordiske ordbøger undersøges, og i Petersson & Sköldberg (2020), hvor beskrivelsen af nogle udvalgte kontroversielle ord fremlægges og diskuteres, herunder *hora* ('hore'), *rödskinn* ('rødhud') og *bög* ('bøsse'). Müller-Spitzer & Rüdiger (2022) fremlægger en undersøgelse af stereotyper i tre genremæssigt forskellige korpora, nemlig et bestående af fiktionstekster, et bestående af avistekster og et bestående af ugeblade/magasiner; en undersøgelse der viser at kønsstereotyper er særligt udbredte i avistekster. Også uden for det leksikografiske område diskuteres problemet med bias og stereotyper i tekster; således konstaterer Huyssteen & Tiberius (2023) at den nyeste udvikling inden for AI og fremkomsten af sprogmodeller som fx OpenAI's GPT-4 og ChatGPT har gjort emnet højaktuelt. Håndteringen af stereotyper i store datamængder medfører i det hele taget mange etiske dilemmaer for firmaer bag sprogteknologiske produkter. Google har valgt at påtage sig et socialt ansvar ved ikke at reproducere bias og stereotyper og beskriver dette som et af deres vigtigste principper, idet de samtidig erkender at det ikke er nemt at afgøre hvad der er stereotypt og kontroversielt, og at det varierer afhængigt af kultur og samfund, se Google's AI Principles, princip 2. Huyssteen & Tiberius (2023) påpeger at problemet ikke kun er reproduktionen af bias og statistisk baserede stereotyper, men at der også i automa-

tisk sprogbehandling og NLP er et behov for at indkredse hvilke ord der i det hele taget er tabubelagte og kontroversielle. Lister over kontroversielle ord er påkrævet til fx automatisk identifikation af krænkende sprog på sociale medier. Oplysning om hvorvidt ord bruges bevidst for at vække anstød eller måske utilsigtet kan opfattes krænkende fordi det er tabubelagt i dele af sprogsamfundet, er et nødvendigt supplement til de polaritetsoplysninger i sentimentleksikoner der anvendes til sentimentanalyse (automatisk genkendelse af negativt, henholdsvis positivt ladede ord). *Det Danske Sentimentleksikon* (Nimb et al. 2022) indeholder også kun polaritetsværdier og ikke oplysninger om hvor kontroversielle ordene er, fx er *gebyrgrib*, *miljøsvin*, *negermusik* og *perker* beskrevet helt ens (stærkeste negative værdi). Information om hvor kontroversielle ordene er (*negermusik* og *perker* er i høj grad, *miljøsvin* og *gebyrgrib* er ikke), ville kunne tilføjes automatisk når DDO-lemmaer er opmærket som beskrevet i denne artikel, idet ord i sentimentleksikonet er koblet direkte til DDO på betydningsniveau. Huyssteen & Tiberius (2023) nævner to eksempler på lister over kontroversielle ord der anvendes i NLP, et italiensk og et japansk.

Der var allerede visse indvendinger mod at anse korpuslingvistiske metoder som absolut objektive i 1990'erne og årene omkring år 2000. Hidalgo-Tenorio (2000) nævner fx at det jo stadig er redaktører der ud af mange forekomster i et korpus udvælger netop de belæg på en betydning som vedkommende anser som bedst egnede til at illustrere den normale sprogbrug. I redaktionen bag den sydafrikanske ordbog *Woordeboek van die Afrikaanse Taal* (WAT, the Bureau of the WAT[1]) baserede man midt i 1990'erne det leksikografiske arbejde på andre principper end CoBuild- og DDO-projekternes når det kontroversielle ordforråd skulle beskrives (Harteveld & van Niekerk 1995 og 1996). The Bureau of

1 Bureau of the WAT (journals.co.za/publisher/botw) er et leksikografisk institut beliggende i Stellenbosch og grundlagt i 1926. Dets hovedopgave er udarbejdelsen af en omfattende betydningsordbog for afrikaans, *Woordeboek van die Afrikaanse Taal*.

the WAT så det som en samfundspligt i tiden efter apartheids ophør i Sydafrika at undlade at reproducere stereotyper. De redaktionsregler der blev formuleret til WAT ud fra denne holdning, er efter min mening blevet meget aktuelle i nutidens leksikografiske arbejde. The Bureau of the WAT udtrykte et direkte ønske om at spille en aktiv rolle i arbejdet med at fremme ligestilling i Sydafrika og dermed udvise forståelse for "a problem which caused great pain, indignation and interpersonal alienation in South Africa" (Harteveld & van Niekerk 1996:393). Hverken i den digitale eller trykte udgave af ordbogen måtte der ved "Insulting and Sensitive Lexical Items" – sensitive eller kontroversielle ord – optræde kollokationer, redaktionelle eksempler eller citater der afspejlede en negativ holdning til befolkningsgrupper. Et eksempel som *you cannot trust a black person with the building process* var fx ikke acceptabelt. I den trykte udgave af ordbogen var man endnu mere restriktiv. Der var hverken kollokationer, redaktionelle eksempler, antonymer, referencer, citater eller andre brugseksempler ved kontroversielle lemmaer, og der måtte ikke bringes krænkende synonymer ("hurtful synonyms"). Racistiske lemmaer blev decideret udeladt i den trykte udgave af ordbogen.

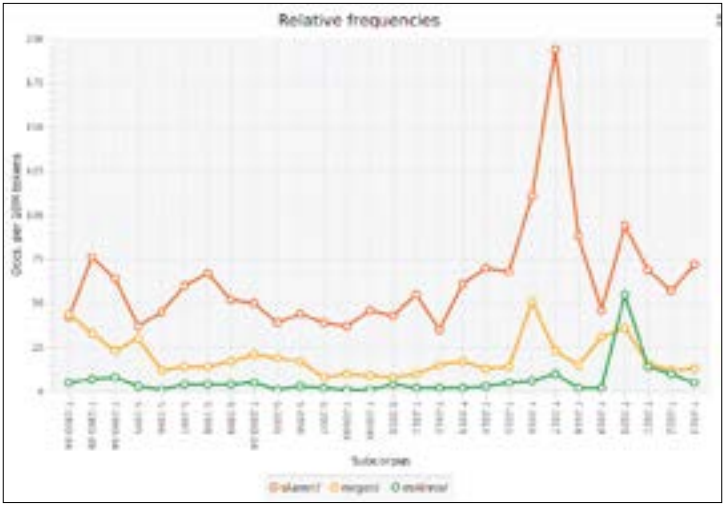
4. Kontroversielle ord: problemstillinger og mulige strategier

I dag har ordbogsredaktioner efter min mening pligt til at forholde sig til problemet med bias og stereotyper i store tekstmængder når ordbogsarbejdet er korpusbaseret. I DDO-projektet arbejdes der fx målrettet med at nedgradere synligheden af ældre stereotype citater fra det oprindelige DDO-korpus og erstatte dem med nyere, mere neutrale citater i den direkte visning (jf. Jensen et al. 2018, Trap-Jensen 2020). Nye redaktionsregler for valørangivelser ved DDO-lemmaer er også under udarbejdelse, bl.a. baseret på erfa-

ringer med redigering af et mindre antal kontroversielle ord relateret til både etnicitet, seksualitet og køn.

I arbejdet med kontroversielle ord indgår to problemstillinger. Den første er selve afgrænsningen af det relevante ordforråd. Petersson & Sköldberg (2020) påpeger at en del af problematikken vedrørende kontroversielle og diskriminerende ord er at de er svære at indkredse. Forfatterne påpeger også at man ikke kan nøjes med at se på de ord der allerede er beskrevet som nedsættende i ordbøger, fordi sproget forandrer sig hurtigt, og fordi ord der hidtil har været neutrale, pludselig bliver diskriminerende eller kontroversielle, fx på grund af påvirkning fra andre sprogsamfund. Fordi vi ved at kontroversielle ord, måske især de ord der ikke er enighed om i sprogsamfundet, typisk debatteres i medierne, kunne et særligt mønster i korpusfrekvens over tid, en såkaldt ordprofil, måske afsløre de ord der pludselig bliver kontroversielle. I perioden hvor et kontroversielt ord debatteres, vil dets frekvens være stærkt stigende i et aviskorpus, herefter vil det formentlig få en lavere frekvens end det havde før mediedebatten startede. Ved hjælp af DDO-redaktionens korpusværktøjer kan dette undersøges i DDO's korpus Bakspejlet (kun delvis offentligt tilgængeligt; bestående af mere end 1,3 milliarder ord fra primært aviser, se også Appel, Sørensen & Jensen, dette bind). Der er en tendens til at hypotesen er korrekt når man ser på ordene *neger*, *slaver* og *eskimo*, som vi med sikkerhed ved har været debatteret offentligt, og hvor i hvert fald *eskimo* som noget relativt nyt nu af mange opfattes som politisk ukorrekt (se figur 1). Men da også ikke-kontroversielle ord af andre grunde vil kunne udvise samme ordprofil i et korpus, kan metoden kun bruges til at få bekræftet en mistanke om at et kontroversielt ord rent faktisk også *er* blevet debatteret på et givet tidspunkt.

Brugerhenvendelser er til gengæld en direkte kilde til indkredsning af potentielt kontroversielle ord. Under den igangværende fjerde bølge af feminisme er det DDO-redaktionens erfaring



Figur 1: Ordprofiler for strengene *slaver*, *neger* og *eskimo* 1930-2023 i korpusset Bakspejlet.

at mange brugere vælger at henvende sig til redaktionen som en form for aktivistisk handling, i tråd med hvad Laura Na Blankholm skriver i dette indlæg: "I den fjerde bølge er internettet og sociale medier en central kampplads" (dagbladet Information 6. oktober 2017). Et godt eksempel på dette er mange samtidige henvendelser til redaktionen i 2019 vedrørende lemmaet *hudfarve*. En bruger skrev fx: "Hvordan kan definitionen på ordet 'hudfarvet' være ferskenfarvet/nordeuropæisk, når der er mange, mange mennesker som ikke er lyse i huden (også her i Danmark)". Et andet eksempel er denne henvendelse vedrørende ordet *grønlænderstiv*; igen blot en af mange samtidige i 2020: "Det er fuldstændig lige meget om ordet findes eller ej. I er med til og dele ordet ud til verden mens vi kæmper her for at få det fjernet! Fjern det fra jeres ordbog." Men selvom brugerhenvendelser bidrager med oplysninger om nye kontroversielle ord, ville det være langt at foretrække at være på forkant, så svaret oftere kunne være at vi allerede har observeret problemet og har noteret ved opslagsordet at beskrivelsen eventuelt bør ændres.

Den anden problematik er leksikografens vurdering af om lemmaet anvendes bevidst nedsættende eller måske kan opfattes som politisk ukorrekt eller krænkende af dele af sprogsamfundet uden at det var tilsigtet – og i hvilken grad det kan. Petersson & Sköldberg (2020) påpeger at det her er næsten umuligt ikke at basere vurderingen på egen sprogfornemmelse, også selv om man undersøger andet materiale som fx slangordbøger, og at det derfor ikke er uden risiko. Brugerhenvendelser til DDO's redaktion viser med al tydelighed at der er stor variation i holdninger til ord inden for køn, ligestilling, seksualitet og etnicitet, en variation der naturligvis også må eksistere blandt leksikografer. Et meget illustrerende eksempel for problematikken er lemmaet *neger*s beskrivelse i DDO. I øjeblikket har ordet sprogbrugsmarkeringen 'oftest nedsættende' for at afspejle at der ifølge redaktørernes opfattelse er tale om divergerende holdninger i sprogsamfundet. Om det evt. er aldersbetinget, er ikke specificeret; den præcise viden om dette kan ikke udledes af korpus. Sprogbrugsoplysningen har ført til to næsten samtidige brugerhenvendelser til DDO-redaktionen. Den ene bruger skriver: "[O]m ordet neger brugt om sorte og afrikanere, skriver i: "oftest nedsættende", men i dagens sprogbrug er det altid nedsættende, medmindre man er meget uoplyst". Den anden bruger skriver: "Jeg ser under opslaget om ordet 'neger', at sprogbrugen ifølge jer oftest er nedsættende. Dette undrer mig, for jeg har aldrig selv brugt ordet nedsættende eller hørt andre bruge ordet nedsættende. Det er et helt almindeligt objektivt ord, som beskriver det, det gør ifølge jeres egen ordbog. En afrikaner. En sort". Med andre ord tilfredsstiller sprogbrugsmarkøren 'oftest nedsættende' tydeligvis hverken den ene eller den anden gruppe. Her står man som redaktør i en vanskelig situation. Det er en nærmest umulig opgave dels at definere mere teoretisk hvordan et tekstkorpus skal opbygges, så det fyldestgørende repræsenterer de forskellige holdninger, dels i praksis at opnå rettigheder til at indsamle tekster til et sådant korpus, fx fra sociale medier. For at

sikre en videnskabelig metode foreslår jeg at man i stedet på en systematisk måde indsamler oplysninger om holdninger blandt flere annotører som – meget vigtigt for metoden – er af forskellig alder og køn, og lader den opnåede viden danne udgangspunkt for den leksikografiske formidling. Jeg vil i afsnit 6 give et bud på hvordan det kan gribes an, men først vil jeg i næste afsnit beskrive en metode der angår den første problematik, nemlig indkredsning af selve ordforrådet.

5. Udpegning af kontroversielle lemmaer

Metoden tager udgangspunkt i ord der allerede er beskrevet som nedsættende i andre ordbøger, i mit eksempel DDO. De lemmaer der er markeret som nedsættende i DDO, udgør ca. 1,5 % af ordbogen, og interessant nok er denne procentdel stabil over tid; den gælder både for det ordforråd der blev redigeret til den trykte DDO i årene 1994-2004, og for det der er redigeret i årene 2007-2024, og som løbende publiceres i DDO online. Valørangivelserne er 'nedsættende', 'nedsættende eller spøgende', 'ofte nedsættende', 'især nedsættende' og 'stærkt nedsættende'. Langt fra alle nedsættende ord i DDO er det vi forstår ved kontroversielle. *Lov om Ligebehandlingsnævnet* nævner i § 1 de områder som nævnet behandler: "Ligebehandlingsnævnet behandler klager over forskelsbehandling på grund af køn, race, hudfarve, religion eller tro, politisk anskuelse, seksuel orientering, alder, handicap eller national, social eller etnisk oprindelse"[2] (2016). Ud fra dette (og idet der ses bort fra ord vedrørende politisk anskuelse) har vi indledningsvis grupperet alle nedsættende ord i den trykte DDO og opstillet fire overordnede kontroversielle kategorier. Den første er 'køn, seksu-

2 Petersson & Sköldberg (2020) anvender i øvrigt, uafhængigt af os, en helt parallel metode på baggrund af lov om diskrimination i Sverige fra 2009 og lignende information på hjemmesiden for den svenske ligestillingsombudsmand.

alitet', den anden er 'handicap, udseende, alder'. Derudover har vi kategorien 'etnisk oprindelse, race, hudfarve, religion/tro' og endelig kategorien 'social oprindelse' (en kategori som vi dog endnu ikke har set så meget på).

Fordi man, som Petersson & Sköldberg (2020) fastslår, ikke kan nøjes med kun at se på ord der allerede er beskrevet som nedsættende, skal listen udvides med de nærsynonymer der ifølge DDO er neutrale, men som har potentiale til at blive nedsættende baseret på deres betydning og ordets eller orddelenes karakter. Da DDO's ordforråd er emneinddelt i *Den Danske Begrebsordbog* (Nimb et al. 2014) og koblet direkte sammen med den på betydningsniveau, kan bruttolisten af kontroversielle ord, både de i forvejen nedsættende og deres nærsynonymer, forholdsvis nemt indsamles på en liste (samme metode anvendtes til at udarbejde *Det Danske Sentimentleksikon* (Nimb et al. 2022)). Begrebsordbogen indeholder 888 navngivne afsnit, herunder også afsnit som 'Fremmede, udlandet', 'Ligestilling', 'Sygdom', 'Sex og begær'. Fordi oplysningen om nedsættende valør er overført automatisk fra DDO til Begrebsordbogens ordforråd (i forenklet form 'neds.'), kan vi nemt fokusere på de grupper i et afsnit der har mange nedsættende ord, se figur 2. Både de i forvejen nedsættende ord og de ord i samme gruppe der vurderes til også at være eller kunne blive kontroversielle og diskriminerende, medtages på listen.

Ved hjælp af denne metode har vi foreløbig indkredset 1.200 kontroversielle og potentielt kontroversielle ord og udtryk, flest af typen 'køn, seksualitet' og 'etnisk oprindelse, hudfarve, national oprindelse' (ca. 400 ord i hver kategori), men også mange vedrørende 'handicap, udseende, alder' (ca. 350 ord, fx *krykhusar*, *klumpfodet*, *dværg*, *hele den pukkelryggede (familie)*; *friluftsgebis*, *hængepatter*, *blåøjethed*, *brunette*; *aldersbyrde*, *gammelmandsbarn*, *bessemor*). Gruppen der vedrører 'social oprindelse', indeholder indtil videre kun 60 ord, fx *bonderøv*, *andedam*, *provinsiel*; gruppen skal suppleres med værdiladede ord der vedrører både lav og

Figur 2: Fire forskellige afsnit der er inddraget fra Begrebsordbogen i indkredsningen af kontroversielle ord: 'Menneskets udseende', 'Indbygger, befolkning', 'Fremmede, udlandet' og 'Tyk'. Nogle af ordene har i forvejen nedsættende valør i DDO (fx *neger* og *flæskebjerg*), andre har ikke (fx *farvet* og *fedtklump*), men bør vurderes og evt. revideres fordi de tilhører samme semantiske felt.

høj social status. Ord der vedrører religion, er ikke endeligt indkredset. Alle fire hovedkategorier bliver løbende udvidet med ord når nye tilfælde opdages, fx på grund af brugerhenvendelser.

Med henblik på at opnå mindre grupper af meget beslægtede ord underinddeles de fire hovedkategorier yderligere. De 400 ord i gruppen 'etnisk oprindelse, hudfarve, national oprindelse' inddeles fx i disse otte underkategorier:

1. ord der afspejler forældet teori om menneskeracer (*mulat*, *halvblods*, *negroid*)
2. ord der afspejler at ikke-vestlig kultur anskues fra et vestligt synspunkt (*naturfolk*, *indianerhyl*)
3. (fremmed) person der er beskrevet ud fra medfødt udseende eller spisevaner (el. anden kropslig egenskab) (*spaghetti* ('italiener'), *den hvide mand*)
4. fremmed folk/folkeslag der er beskrevet med en ikke-national/ikke-geografisk betegnelse (*kaffer*, *sigøjner*, *lap*/*laplænder*)
5. ord hvor etnicitet er del af et (negativt) ord eller udtryk med anden betydning (*grønlænderstiv*, *negerbolle*, *bande som en tyrk*, *du milde kineser*, *fransk horeunge*, *squaw* (i den overførte betydning 'hustru'))
6. ord vedrørende indvandring i Danmark (*perker*, *perkerdansk*)
7. ord vedrørende kolonitiden (*koloniherre*, *slaveopstand*)
8. evt. værdiladet, ikke-officiel betegnelse for statsborger (*thaipige*, *kineserinde*, *kartoffeltysker*)

På denne måde kan tæt beslægtede beskrivelser i DDO nemmere sammenlignes ud fra en diskriminationsvinkel og en ligestillingsbetragtning. Man ser fx hurtigt at nogle af ordene ikke har nogen sprogbrugsoplysning i DDO, selv om de i dag kan opfattes nedsættende eller krænkende. Nogle eksempler er *burkabil* og *squaw*. Ligeledes har *gringo*, *lap* og *laplænder* samt udtrykket *bande som en tyrk* ('bande kraftigt og ofte') ingen valørangivelse. Substantivet *lapp* er både markeret som stærkt nedsættende og gammeldags i *Svensk Ordbok*, og den danske ækvivalent bør derfor undersøges nærmere. Man opdager også at der kan være forskel på valøroplysninger ved ord der i øvrigt ligner hinanden, fx er *flæskebjerg* ('overvægtig person') 'nedsættende', mens *fedtklump* i samme betydning blot er 'uformelt'. Og *abekat* ('person der opfører sig vildt

og uciviliseret eller tåbeligt og naragtigt'), er 'nedsættende eller spøgende', mens *barbar* ('rå person der stammer fra et fremmed, uciviliseret land') blot er markeret 'især historisk'.

Nyligt indsatte, mere detaljerede valøroplysninger ved prøveord i DDO kan også være meget forskellige inden for undergrupperne (som nævnt arbejdes der i disse år med at opstille nye redaktionsregler). I tabel 1 ses eksempler på sprogbrugsoplysninger for underkategorien 'ord der afspejler at ikke-vestlig kultur anskues fra et vestligt synspunkt':

heksedoktor	person der menes at besidde magisk-religiøse kræfter til bl.a. at helbrede sygdomme, beskytte mod fremmed magi og spå om fremtiden – kendes især fra stammesamfund SPROGBRUG forældet | nu oftest nedsættende
indianer	person tilhørende et af de folk som sammen med inuit og aleuter udgør den oprindelige befolkning i Nord- og Sydamerika – som alternativ bruges nu ofte udtrykket *oprindelig/indfødt amerikaner*, idet *indianer* kan opfattes stødende
indianerblod	**have indianerblod i årerne**: tilhøre eller være efterkommer af den oprindelige befolkning i Amerika (med undtagelse af det nordligste Nordamerika) SPROGBRUG kan virke stødende
indianerhyl	meget kraftigt og vildt hyl SPROGBRUG ordet kan være problematisk fordi det afspejler en nu forældet opfattelse af Amerikas indfødte befolkning som vild og blodtørstig, som fx fremstillet i ældre westernfilm
naturfolk	folk der lever i nær kontakt med og i umiddelbar afhængighed af naturen under anvendelse af enkel teknologi SPROGBRUG denne brug kan være problematisk da den indebærer en forsimplet modstilling af natur og kultur

Tabel 1: Sprogbrugsoplysninger for ord i underkategorien 'ord der afspejler at ikke-vestlig kultur anskues fra et vestligt synspunkt'.

De opstillede lister og undergrupper kan danne udgangspunkt for det videre arbejde frem mod ændrede redaktionsregler. Den store udfordring er at opnå viden om ordenes valør, idet korpusundersøgelser suppleret med introspektion som beskrevet ovenfor

ikke nødvendigvis er repræsentativt for de holdninger der findes i sprogsamfundet.

6. Opmærkning af kontroversielle ord

For at sikre detaljeret viden om holdninger til sensitive ord, herunder den variation der ofte er i holdninger blandt sprogbrugerne, kan man vælge at bede en række lingvister af forskellig alder og køn opmærke ordene. Man kan fx tage udgangspunkt i de oplysningstyper med et tilhørende lukket inventar af værdier der foreslås af Huyssteen & Tiberius (2023). Oplysningstyperne er fastlagt i et samarbejde mellem en sydafrikansk og en hollandsk leksikografisk institution og baseret på en tilbundsgående analyse af problemstillingen med kontroversielle ord. Begge lande har stor erfaring med at arbejde med leksikografiske beskrivelser af det kontroversielle og sensitive ordforråd. I Holland har man fx gennem en lang årrække haft tradition for at udgive særlige ordbøger der beskriver det tabubelagte ordforråd, og situationen i Sydafrika er omtalt ovenfor. I store træk er de nødvendige oplysningstyper ifølge de to forfattere følgende (udfyldt med mine egne intuitive bud på danske eksempler). To oplysninger anvendes til at beskrive hvor ofte og i hvor høj grad ordet er kontroversielt:

- I hvor høj grad er lemmaet tabubelagt (<taboo Value>)? *Nigger* er fx i højeste grad, *thaipige* i lav grad.
- Hvor prototypisk for lemmaet er den kontroversielle betydning (<taboo Prototypicality>)? Hvis ordet uanset kontekst altid er kontroversielt, er svaret 'altid' (fx *nigger* og *neger*), andre værdier er 'ofte' (fx *indianer*), 'undertiden' (fx *slave*) og 'sjældent' (fx *handicappet*).

Hensigt og virkning ved anvendelse af ordene undersøges desuden ved hjælp af tre oplysningstyper:

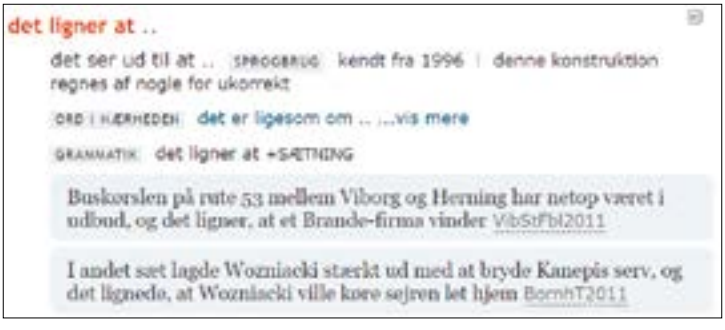
- Bruges lemmaet (i denne betydning) ofte eller mest som skældsord (<speechAct>)? I tilfældene *bøsserøv* og *nigger* er svaret ja.
- I hvor høj grad har den talende en bevidst hensigt med at bruge krænkende sprog når lemmaet i denne betydning anvendes <illocution>? *Sortsmudsker*, *negermusik* og *betonlebbe* er fx bevidst nedsættende.
- Opfatter modtageren dette lemma som stærkt krænkende, lidt krænkende, politisk ukorrekt eller ingen af delene (<perlocution>)? *Indianer* og *uland* opfattes fx ofte som politisk ukorrekt, *mulatpige* som krænkende.

Endelig er der mulighed for at angive et eventuelt neutralt synonym (<orthophemism>), fx *roma* som alternativ til *sigøjner*.

Også ord der ikke umiddelbart er kontroversielle, men som har potentiale til at blive det, fx fordi første- eller sidsteleddet er kontroversielt, skal opmærkes af lingvisterne og vil formentlig få så lave værdier at de blot sættes på en observationsliste. Et eksempel kunne være substantivet *slaverom* på grund af førsteleddet *slave-* og sprogbetegnelsen *eskimoisk* på grund af *eskimo*s kontroversielle karakter.

Idealet er at yngre og ældre sprogbrugere af forskelligt køn er repræsenteret blandt annotørerne, og at det blot er deres intuitive holdninger til de enkelte ord der skal anføres, så arbejdet ikke er for tidskrævende. Når intuitive holdninger indsamles blandt en større og til en vis grad repræsentativ gruppe personer, sikrer man bedre at variationen i holdninger i sprogsamfundet afspejles (jf. Ipsos MORI (2021), hvor fokusgrupper sammensat af personer af forskelligt køn, alder og etnicitet bl.a. danner udgangspunkt for sproglige holdningsanalyser). De ord der får varierende oplys-

ninger på tværs af annotørgruppen, er selvfølgelig særligt udfordrende at beskrive i en ordbog. Man kan vælge at lade variationen udgøre en del af beskrivelsen ud fra den betragtning at det udgør en vigtig viden om ordet der bør videreformidles. Plank (2022) argumenterer fx for at variation blandt sprogbrugere og annotører er meget vigtig information der bør inkluderes i annoterede data frem for den hidtidige praksis i NLP (og inden for leksikografien) hvor lingvister og leksikografer forhandler sig til enighed eller beslutter sig for kun at anvende én af værdierne. I NLP opererer man traditionelt med såkaldte guldstandarder med kun én værdi som datagrundlag. I leksikografi kan oplysninger altid modificeres (fx i form af 'ofte(st)' i sprogbrugsmarkøren 'oftest nedsættende' ved *neger* i DDO), men det fremgår sjældent tydeligt hvori variationen består, og det kan, som beskrevet ovenfor, skabe problemer. Variation med hensyn til grammatisk korrekthed fremgår fx mere tydeligt i DDO. Ved udtrykket *det ligner at* .. anføres det fx at "denne konstruktion regnes af nogle for ukorrekt", se figur 3. Her er det underforstået at *nogle* refererer til personer der er uddannet inden for danskfaget.



Figur 3: Udtrykket *det ligner at* .. har i DDO oplysningen: "denne konstruktion regnes af nogle for ukorrekt".

Man kan også vælge en mere enkel løsning, fx blot at angive en advarsel ved ordet eller at tolke krænkende virkning som et ud-

slag af at brugen i dag er blevet 'gammeldags', vel at mærke i de tilfælde hvor det er de ældre annotører der opfatter ordet som uproblematisk, mens de yngre annotører undgår at bruge ordet og opfatter det som krænkende (i den pågældende betydning). *Bokmålsordboka* angiver fx blot at *eskimo* er en "foreldet betegnelse", og *Svensk Ordbok* angiver blot at *flicka* i betydningen 'kvinna' er "något ålderdomligt". Hvis det er svært at udlede ud fra korpusundersøgelser (idet skribenternes alder ikke kendes), bør grundlaget for denne formidling ideelt set også bygge på indsamlede data fra en aldersmæssigt bredt sammensat gruppe. I princippet bør opmærkningsarbejdet også gentages forholdsvis ofte sammenlignet med andet nødvendigt revideringsarbejde i et ordbogsprojekt, fx udgiver BBC (BBC blog (2010)) opdaterede sproglige vejledninger til medarbejderne vedrørende kontroversielle ord i deres udsendelser hvert fjerde-femte år. Det bør derfor udformes som en ikke alt for tidskrævende annotationsopgave, hvor alene intuition danner grundlag for den enkelte annotørs opmærkning, men hvor antallet af annotører og spredning i alder og køn derimod er meget afgørende.

7. Konklusion

Der er mange leksikografiske udfordringer i arbejdet med at undersøge og beskrive kontroversielle ord i en erkendelse af at de tekstkorpora, der i øvrigt danner et rigtig godt udgangspunkt for deskriptivt leksikografisk arbejde, per se er biased og indeholder mange stereotyper. Problematikken er højaktuel på grund af den øgede opmærksomhed på ligestilling i samfundet og fremkomsten af AI og sprogmodeller der bygger på statistiske beregninger på store tekstmængder. I det leksikografiske arbejde er stereotype bigrammer og citater ud fra vores erfaring nemmest at håndtere redaktionelt, hvorimod angivelsen af sprogbrugsmarkører

ved kontroversielle ord indebærer flere udfordringer. Ordene er svære at indkredse i ordbog og korpus, og det er svært at tildele sprogbrugsmarkører fordi der kan være stor variation i holdningen til ordene blandt både sprogbrugere og ordbogsredaktører. Leksikografer har ekspertisen til at indsamle og beskrive det sensitive ordforråd i detaljer, og man kan med fordel anvende annotationsmetoder der kendes fra arbejdet med sprogdata i NLP. Man opnår et videnskabeligt datagrundlag hvis både ældre og yngre informanter af forskelligt køn involveres i opmærkning af ordforrådet, og det kan også anvendes til at formidle interessant viden om variation blandt sprogbrugerne. For at være på forkant med sprogudviklingen i samfundet bør også ord der blot har potentiale til at blive kontroversielle, opmærkes. Udfordringerne er at annotering er tidskrævende, især når mange skal involveres, og at det hurtigt forældes fordi sprogbrugen måske særligt i disse år er i hastig udvikling inden for området.

Litteratur

Ordbøger, korpusser og digitale resurser

Bakspejlet = DDO's interne korpus, Det Danske Sprog og Litteraturselskab (2024).

BBC blog (2010): New edition of BBC's Editorial Guidelines. <bbc.co.uk/blogs/theeditors/2010/10/new_edition_of_bbc_editorial_g.html> (archived page).

Begrebsordbogen = *Den Danske Begrebsordbog*.

Bokmålsordboka. Språkrådet og Universitetet i Bergen. <ordbokene.no/nob/bm> (april 2024).

Den Danske Begrebsordbog (2014). Sanni Nimb, Henrik Lorentzen, Liisa Theilgaard & Thomas Troelsgård. København/Odense: Det Danske Sprog- og Litteraturselskab og Syddansk Universitetsforlag.

Den Danske Ordbogs korpus = Ole Norling-Christensen & Jørg Asmussen (1998): The Corpus of The Danish Dictionary. I: *Lexikos* 8. doi:10.5788/8-1-955.

COBUILD (1987) = *Collins COBUILD English Language Dictionary*. Editor in Chief: John Sinclair, Managing Editor: Patrick Hanks. London/Glasgow: Collins.

DDO (2003-05) = *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab og Gyldendal.

DDO online = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ddo> (april 2024).

Det Danske Sentimentleksikon = <github.com/dsldk/danish-sentiment-lexicon> (april 2024).

Google's AI Principles = <ai.google/responsibility/principles> (april 2024).

Lov om Ligebehandlingsnævnet (2016, LBK nr. 1230 af 02/10/2016). Bekendtgørelse af lov om Ligebehandlingsnævnet. Beskæftigelsesministeriet. <retsinformation.dk/eli/lta/2016/1230>.

Moths Ordbog. Det Danske Sprog- og Litteraturselskab. <mothsordbog.dk> (april 2024).

Ordbog over det danske Sprog. Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ods> (april 2024).

Svensk Ordbok = *Svensk Ordbok utgiven av Svenska Akademien*, Göteborgs universitet. <gu.se/svenska-spraket/svensk-ordbok> (april 2024).

Woordeboek van die Afrikaanse Taal (WAT). <woordeboek.co.za> (april 2024).

Anden litteratur

Appel, Kirsten, Nathalie Hau Sørensen & Jonas Jensen (2024): Jagten på hverdagssproget – brugen af tekster fra internetfora i arbejdet med Den Danske Ordbog. I: *LexicoNordica* 31 (dette bind).

Fjeld, Ruth Vatvedt (2015): Om ordbokseksempler og stereotypisering av kjønn i noen nordiske ordbøker. I: Caroline Sandström, Ilse Cantell, Eija-Riitta Grönros, Pirkko Niolijärvi & Eivor Sommadahl (red.): *Perspektiv på lexikografi, grammatik och språkpolitik i Norden*. Helsingfors: Institutet för de inhemska språken. 35-65.

Harteveld, Pieter & Angelique E. van Niekerk (1995): Policy for the Treatment of Insulting and Sensitive Lexical Items in the *Woordeboek van die Afrikaanse Taal*. I: *Lexikos* 5. doi:10.5788/5-1-1068.

Harteveld, Pieter & Angelique E. van Niekerk (1996): Policy for the Treatment of Insulting and sensitive Lexical Items in the Woordeboek van die Afrikaanse Taal. I: Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström & Catarina Röjder Papmehl (red.): *Euralex '96 Proceedings I-II*. Gøteborg: Göteborg University, Department of Swedish. Part 1, 381-393.

Hidalgo-Tenorio, Encarnación (2000): Gender, Sex and Stereotyping in the Collins COBUILD English Language Dictionary. I: *Australian Journal of Linguistics 20(2)*, 211-230. doi:10.1080/07268600020006076.

Huyssteen, Gerhard B. van & Carole Tiberius (2023): Towards a lexical database of Dutch taboo language. I: Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček & Simon Krek (eds.): *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27-29 June 2023*. Brno: Lexical Computing CZ s.r.o. 53-74. <elex.link/elex2023/wp-content/uploads/elex2023_proceedings.pdf>.

Ipsos MORI (2021): *Public attitudes towards offensive language on TV and radio: Summary Report*. <ipsos.com/en-uk/public-attitudes-towards-offensive-language-tv-and-radio> (april 2024).

Jensen, Jonas, Henrik Lorentzen, Sanni Nimb, Mette-Marie Møller Svendsen & Lars Trap-Jensen (2018): Thaipiger, muskelhunde og fulde svenskere: nedsættende ord, stereotyper og ligestilling i *Den Danske Ordbog*. I: Ásta Svavarsdóttir, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdís Úlfarsdóttir (red.): *Nordiske Studier i Leksikografi* 14. Reykjavik: Nordisk Forening for Leksikografi. 141-151.

Müller-Spitzer, Carolin & Jan Oliver Rüdiger (2022): The influence of the corpus on the representation of gender stereotypes in the dictionary. A case study of corpus-based dictionaries of German. I: Annette Klosa-Kückelhaus, Stefan Engelberg, Christine Möhrs & Petra Storjohann (eds.): *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany*. Mannheim: IDS-Verlag. 129-141. <euralex.org/wp-content/uploads/2022/09/EURALEX2022_Proceedings.pdf>.

Nimb, Sanni, Nikolai Hartvig Sørensen & Thomas Troelsgård (2018): From Standalone Thesaurus to Integrated Related Words in The Danish Dictionary. I: Jaka Čibej, Vojko Gorjanc, Iztok Kosem & Simon Krek (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts. 916-923. <euralex.org/publications/from-standalone-thesaurus-to-integrated-related-words-in-the-danish-dictionary>.

Nimb, Sanni, Sussi Olsen, Bolette Pedersen & Thomas Troelsgård (2022): A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. I: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, Stelios Piperidis (eds.): *Proceedings of the Thirteenth Language Resources and Evaluation Conference: LREC2022*. Marseille: European Language Resources Association. 2826-2832. <aclanthology.org/2022.lrec-1.302>.

Petersson, Stellan & Emma Sköldberg (2020): To discriminate between discrimination and inclusion: a lexicographer's dilemma. I: Zoe Gavriilidou, Maria Mitsiaki & Asimakis Fliatouras (eds.): *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress*. Alexandroupolis: Democritus University of Thrace. 381-386. <euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_ProceedingsBook-Vol1.pdf>.

Plank, Barbara (2022): The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. I: Yoav Goldberg, Zornitsa Kozareva & Yue Zhang (eds.): *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. EMNLP 2022*. Abu Dhabi: Association for Computational Linguistics. <aclanthology.org/2022.emnlp-main.731.pdf>.

Scheuer, Jann (1995): Hans hustru og hendes bryster. I: Mette Kunøe & Erik Vive Larsen: *5. Møde om Udforskningen af Dansk Sprog. Aarhus Universitet 13.-14. oktober 1994* (MUDS 5). Aarhus: Aarhus Universitetsforlag. 246-256.

Trap-Jensen, Lars (2020): Inklusion eller mindretalsdiktatur? Om politisk korrekthed, minoritetshensyn og leksikografisk deskriptivisme i *Den Danske Ordbog*. I: *LexicoNordica* 27, 137-156.

Williams, Geoffrey (2003): From meaning to words and back: Corpus linguistics and specialised lexicography. I: ASp 39-40 Groupe d'étude et de recherche en anglais de spécialité. 91-106. doi:10.4000/asp.1320.

Sanni Nimb
seniorredaktør, ph.d.
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
sn@dsl.dk