Korpus eller brugerne – hvem får det sidste ord?

Af *Lars Trap-Jensen*

Siden årtusindskiftet har ordbogsverdenen gennemlevet en revolution for at omstille sig til den digitale verdens vilkår. Ændringerne har været mere omfattende end omverdenen måske er klar over. Det har ikke kun været et spørgsmål om at omlægge produktionen fra papir til skærm, nej selve det at lave ordbøger er i dag grundlæggende anderledes end det var for 15 år siden. Noget af det der blandt andet har ændret sig, er muligheden for at følge brugerens faktiske søgeadfærd i en digital ordbog. Spørgsmålet er om den viden i højere grad bør udnyttes til at udvælge de ord som står i ordbogen.

I denne artikel drejer det sig om nye ord til en eksisterende ordbog, nærmere bestemt udvælgelse af nye ord til vedligeholdelse af lemmabestanden i *Den Danske Ordbog* (DDO), og det skal understreges at nye ord her bruges i en lidt bredere forstand end normalt. Et nyt ord er i denne sammenhæng ganske enkelt hvilket som helst ord der ikke tidligere har været i ordbogen. Det behøver altså ikke være en neologisme, men kan også være et eksisterende ord der hidtil er blevet overset eller fravalgt.

DDO er en korpusbaseret ordbog. Bag enhver korpusbaseret ordbog ligger en antagelse om at korpusfrekvens mere eller mindre afspejler et ords udbredelse i sproget, og endvidere at argumentet for at tage et ord med i ordbogen styrkes med stigende korpusfrekvens. Hvis sandheden skal frem, ved vi dog ikke om det også er de ord brugerne faktisk slår op, fordi vi hidtil ikke har haft tilstrækkeligt empirisk undersøgelsesmateriale til at sige ret meget om sammenhængen mellem korpusfrekvens og søgeadfærd. Ældre undersøgelser har typisk haft et forholdsvis spinkelt empirisk grundlag, hos de Schryver & Joffe (2004) 21.337 opslag og hos Bergenholtz & Johnsen (2005) 1.016.960 opslag. Den situation har imidlertid ændret sig idet ordbøgerne efterhånden har været online i nogen tid. I DDO's tilfælde er der tale om mere end fire år, og i den periode er samtlige søgninger i ordbogen blevet registreret og gemt i en søgelog. Den eneste undersøgelse med sammenlignelig empirisk basis som jeg kender til, er Bergenholtz & Norddahl (2012), der har undersøgt logfiler for Den Danske Netordbog over en næsten tilsvarende periode, 31½ måned, og med et månedligt antal opslag med match på 568.062, svarende til ca. 80 % af de tilsvarende opslag i DDO.

Vi besluttede at undersøge hvilken sammenhæng der kan være mellem de ord brugerne faktisk søger efter, og disse ords repræsentation i korpus, dvs. deres frekvens[1]. Nærmere bestemt ville vi undersøge den eksisterende ordbestand i DDO og finde ud af hvilke søgemønstre der var

knyttet til den. Ved at sammenligne søgeloggen med korpusudbredelsen ville vi prøve at få svar på spørgsmål som disse:

(1) Hvilke af DDO's ord slås især op, og hvad karakteriserer dem? Er det hyppige eller sjældne ord; er det især ord fra bestemte ordklasser?

(2) Hvilke andre ord end dem der står i ordbogen, slår brugerne op, og hvad karakteriserer dem?

Hvis svaret på spørgsmål (2) afviger markant fra svaret på (1), kan det give anledning til at revidere principperne for lemmaselektionen fremover. Men først et par ord om selve undersøgelsens anlæg.

1. Undersøgelse af logfiler

1.1. Metodiske valg og forbehold

Vi valgte at bruge det originale DDO-korpus i undersøgelsen (se nærmere herom i Norling-Christensen & Asmussen 1998). Dette korpus udmærker sig ved at have en velafbalanceret fordeling af tekster på genrer, medier og forfatternes sociologiske parametre og sikrer derved at de fundne resultater er statistisk valide. Til gengæld har det også nogle åbenlyse mangler. For det første er korpusset med sine ca. 40 mio. løbende ord ikke vældig stort målt med moderne standarder, og for det andet har det efterhånden nogle år på bagen. Det dækker perioden 1983-1992 og rummer således i sagens natur ikke neologismer der er kommet ind i sproget senere (se også Lorentzen & Nimb 2011). Alligevel måtte hensynet til datas validitet veje tungere end kvantitet og aktualitet.

Et andet forbehold man er nødt til at tage, gælder søgeloggen. Som udgangspunkt viser den kun de søgestrenge brugerne har indtastet i søgefeltet, og man kan ikke altid vide med sikkerhed hvad en bruger har villet slå op. Hvis en indtastet streng fx kan henføres til flere opslagsord, fx i tilfælde af homografi eller hvis der er sammenfald mellem en bøjningsform og et opslagsord, er det blevet talt som et match på begge ord. Dernæst kan søgninger være foretaget af både mennesker og robotter. Robotcrawling kan være nyttigt til forskellige formål, men ikke i dette tilfælde. Vi har kun været interesseret i søgninger foretaget af rigtige mennesker. Derfor har vi bestræbt os på at eliminere Googles crawlere, ondsindede hackerangreb og testsøgninger fra søgeloggen. Dog er det ikke nemt at fjerne lige præcis alle disse tilfælde og ingen andre. Det er muligt at der nogle gange er fjernet for meget, andre gange for lidt. Et tilsvarende forbehold gælder metasøgninger. Når brugerne har skrevet 'synonym' eller 'ordbog' i Googles søgefelt og er blevet henvist til Den Danske Ordbogs opslag for pågældende ord, er det overvejende sandsynligt at de snarere ville finde frem til resursen frem for ordene. Men helt sikker kan man naturligvis aldrig være.

Et eksempel på hvordan søgeloggens registrering af søgninger ser ud, ses i figur 1.

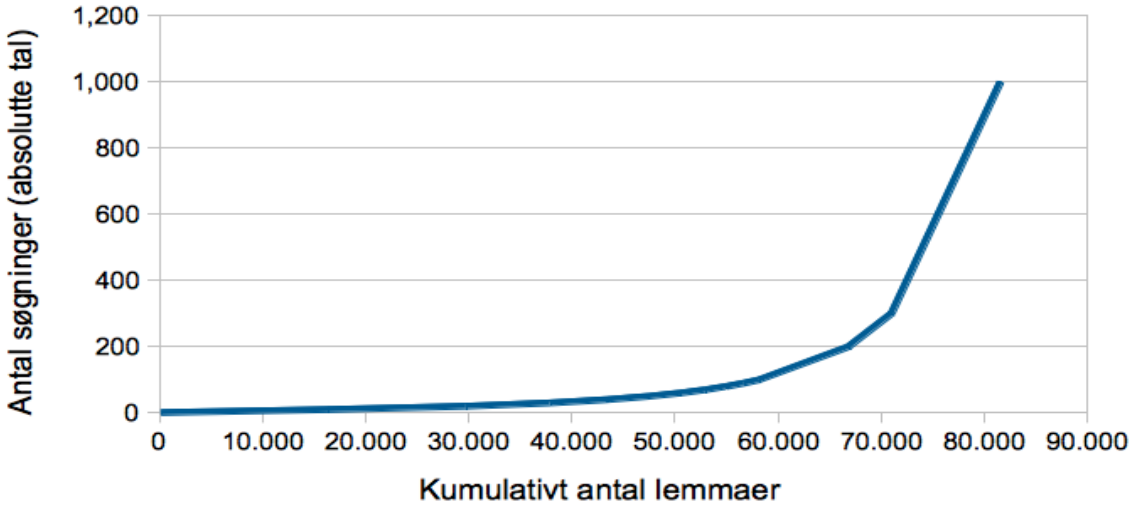
id	query	normalized	caller	ip_address	found	created
228751	rekvisition	rekvisition	query	*A2E6DE5155C2E3E243B692EC03A794C1E26C6700	yes	2009-12-01 01:56:05
228759	forælder	forælder	query	*144363857FF3C34DEBB2C2AECEC0CC05C39D515B	yes	2009-12-01 02:05:21
228763	trapeze	trapeze	query	*0A73B32E298808D0FFA3FD8201B7CF8EF170A7B7	no	2009-12-01 02:07:13
228772	forstillelse	forstillelse	query	*4FDD3D9192DD1DE1F2C831D9971B9BC96347E394	yes	2009-12-01 02:12:27
228777	vanskelig	vanskelig	query	*9716036C15A5DB086E6ECB345D72A6FF1F11CE9B	yes	2009-12-01 02:14:45
228784	skrømt	skrømt	query	*4FDD3D9192DD1DE1F2C831D9971B9BC96347E394	yes	2009-12-01 02:17:08
228785	kareen	kareen	query	*606D60B3C3016918709759B2B44AE7BCCEF68588	no	2009-12-01 02:17:47
228787	kapere	kapere	query	*606D60B3C3016918709759B2B44AE7BCCEF68588	yes	2009-12-01 02:17:54
228788	karreen	karreen	query	*606D60B3C3016918709759B2B44AE7BCCEF68588	no	2009-12-01 02:18:07
228793	værge sig	værge sig	query	*4FDD3D9192DD1DE1F2C831D9971B9BC96347E394	yes	2009-12-01 02:19:13

Figur 1: Eksempel fra søgeloggen

1.2. Undersøgelsens omfang og overordnede resultater

Søgeadfærden blev undersøgt for en treårig periode, fra december 2009 til december 2012. I den periode blev der foretaget i alt 29.551.938 søgninger i Den Danske Ordbog, heraf 2.208.872 forskellige søgninger. Af det samlede antal søgninger var de 24.478.138 vellykkede på den måde at de havde et eksakt match i basen, mens der var 5.073.800 forgæves søgninger, hvoraf de 1.924.615 var forskellige søgninger.

En simpel sammenligning mellem DDO's lemmabestand og søgeloggen er i sig selv interessant at foretage. Den kan måske be- eller afkræfte nogle af de myter som findes blandt leksikografer og ordbogsbrugere, af den enkle grund at vi ikke tidligere har haft viden om hvilke ord brugerne slår op. Én påstand lyder: Der er mange ord i en ordbog som aldrig bliver slået op. Taget helt bogstaveligt er det en påstand som ikke holder: I alt er der blot 202 opslagsord i ordbogen der optræder i søgeloggen med frekvensen 0, altså kun omkring 2 promille af ordbogens i alt knap 100.000 opslagsord. Man kan naturligvis indvende at en søgefrekvens på 0 er meget strikt. Sætter man i stedet grænsen ved 3 søgninger, stiger tallet til at omfatte 5.000 opslagsord, og sættes det til 10, drejer det sig om 16.000 ord. Hvis man fortsætter med at øge kravet til søgefrekvens på tilsvarende måde, kan sammenhængen afbildes grafisk. Det er hvad der er gjort i figur 2. Man kan se at der er mange ord med en forholdsvis lav søgefrekvens, mens der kun er 10-20.000 ord som bliver slået op mange gange. Det kan være en opmuntring til dem der laver ordbøger med et begrænset antal opslagsord. Det ser ud til at de faktisk dækker en ret stor del af brugernes behov.



Figur 2: Sammenhæng mellem søgefrekvens og antal lemmaer

Vi kan altså sige at det er en myte at mange ord aldrig bliver slået op, men der er alligevel noget om snakken, idet mange ord kun bliver slået op vældig lidt. Samtidig gælder det at der er et begrænset antal ord, ca. 20.000, som har en meget høj søgefrekvens.

Resultatet er bemærkelsesværdigt fordi det adskiller sig markant fra Bergenholtz & Norddahl (2012), der i deres undersøgelse når frem til at 66,6 % af alle artikler i Den Danske Netordbog har søgefrekvensen 0. Forskellen er så iøjnefaldende at den ikke kan være tilfældig. De mest nærliggende forklaringer kan være: 1) lemmabestanden i de to ordbøger er væsensforskellig, 2) de to ordbøger har forskellige brugergrupper med forskellige behov og 3) adgangen til ordbøgernes indhold adskiller sig så meget at det fører til forskellig søgeadfærd. Umiddelbart virker den sidste forklaring mest overbevisende: Hvor DDO kan benyttes gratis og brugerne primært kommer via søgninger på Google, kommer de fleste opslag i Den Danske Netordbog via søgninger i betalingstjenesten *ordbogen.com* (i hvert fald ud over de to gratis opslag om dagen). Forklaringen kan dog også være en kombination af ovennævnte faktorer eller indebære andre som ikke er nævnt her. Det fortjener en nærmere undersøgelse.

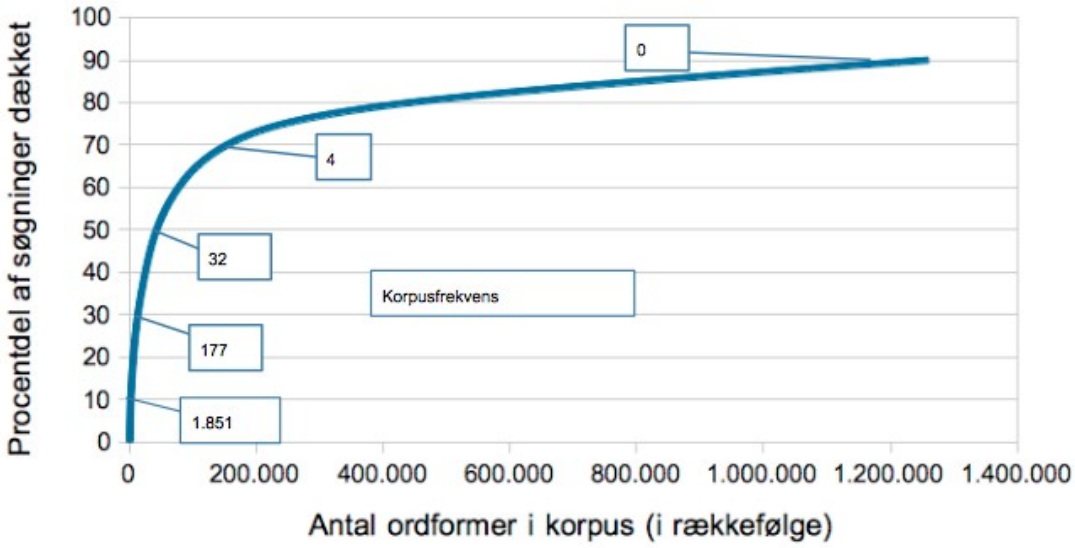
Et forhold som man skal være opmærksom på ved sammenligningen, er det tidsmæssige aspekt. Det er klart at ord der først er blevet tilføjet til ordbogen sent i løbet af den undersøgte periode, ikke har samme chance for at blive søgt som de ord der har været tilgængelige i hele perioden. Ordet *sprogteknolog* hører fx til blandt ordene med søgefrekvens 0, men da dette opslagsord først kom med i forbindelse med en opdatering den 29. november 2012, giver det naturligvis ikke et retvisende billede. Heldigvis er problemet ikke så omfattende, idet antallet af ord der ikke har været til stede i basen gennem hele perioden, begrænser sig til 1.426.

En anden påstand der ofte fremsættes, er at sprogets almindelige og centrale ord sjældent eller aldrig slås op, for det er jo ord vi alle kender og derfor ikke har brug for at slå op. Tallene fra søgeloggen viser imidlertid noget andet, og påstanden kan med overraskende stor sikkerhed afvises som en myte uden hold i virkeligheden. Der er således mange ord fra de lukkede ordklasser, de grammatiske funktionsord, blandt de hyppigst søgte ord: 64 funktionsord blandt de 500 mest søgte ord overhovedet. Det samme gælder sprogets almindelige ord i det hele taget: Blandt de 1000 mest søgte ord er mere end en fjerdedel også mellem de 1000 mest frekvente ord i korpus, og næsten 60 % hører til de 10.000 mest frekvente ord i korpus. Den omvendte tendens gælder også: Blandt de ord der kun er søgt efter 0, 1 eller 2 gange, er der ikke et eneste funktionsord. De sjældnest søgte ord udgøres stort set kun af substantiver og adjektiver, og der er ingen simpleksord imellem. Typiske eksempler på sjældne søgninger er sammensætninger som fx *broderie-anglaise-flæse, middagsradioavis* og *rosenkålssuppe*. Tallene kan måske fortælle noget om hvem der især bruger ordbogen. Vi har dog ikke undersøgt hvordan søgningerne fordeler sig på IP-adresser, altså om tendensen også holder hvis man tager højde for at nogle brugere er meget flittige til at slå op. Fx er lørnere typisk meget flittige brugere der har brug for også at slå funktionsord og andre almindelige ord op. Det kunne være interessant om der er sådan en sammenhæng, men det har vi altså ikke haft mulighed for at undersøge.

2. Søgelog og korpusrepræsentation

Hvis den korpusbaserede lemmaselektion er et sundt princip, må man gå ud fra at de ord brugerne søger efter, også er repræsenteret i korpus – med de forbehold som tidligere blev taget over for bl.a. neologismer. Er det tilfældet, er de nemlig enten allerede optaget i DDO eller må forventes at blive det som led i den løbende opdatering.

En måde at undersøge i hvor høj grad det er tilfældet, er at se på ordformerne i korpus og deres mulighed for at imødekomme de faktiske søgninger. Vi sorterede derfor alle ordformer i korpus i rækkefølge efter faldende frekvens og målte hvor stor en procentdel af søgningerne de var i stand til at dække. Sammenhængen illustreres i figur 3.



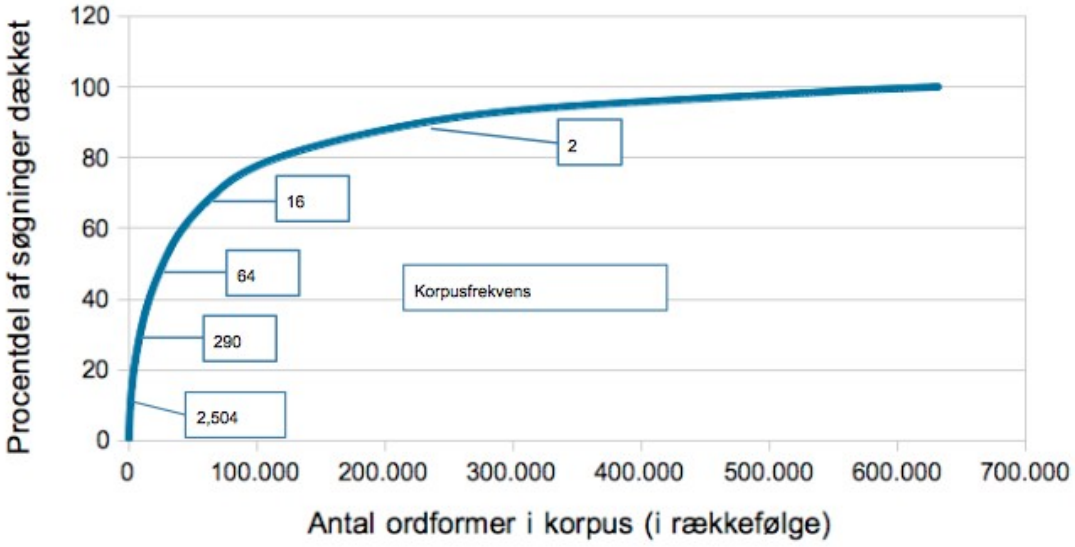
Figur 3: Sammenhæng mellem korpusfrekvens og korpusdækning

Af figuren fremgår det at for at returnere 30 % af opslagene kræves der 12.774 ordformer i korpus. Når ordformerne som her er ordnet efter faldende frekvens, kan vi se at ordform nr. 12.774 har en absolut frekvens i korpus på 177. De absolutte frekvenser for udvalgte procentdele vises i de små bokse i figur 3. Man kan således se at der for at dække halvdelen af søgningerne kræves 42.380 ordformer, og boksen der peger på det sted hvor kurven skærer 50 %-grænsen, angiver at den pågældende ordform forekommer 32 gange i korpus, mens 70 %-grænsen svarer til 4 former i korpus.

Bliver man ved på den måde, ville vi ideelt set gerne kunne returnere alle søgninger, men det viser sig umuligt af to grunde: For det første kræver det flere ordformer end der er i det brugte korpus. For at nå 90 % skulle vi have ca. 1,3 mio. ordformer, men korpus rummer kun omkring 600.000 former. For det andet forekommer der støj i søgestrengene, hvilket dækker over at der dels indtastes mange nonsens-ord, dels er fejlstavninger i søgestrengene som ikke vil optræde i et korpus næsten uanset størrelsen. Det vil derfor nok aldrig blive muligt at nå op på 100 % selvom antallet af ord i korpus blev mangedoblet.

Hvad man i stedet kan gøre, om ikke andet som et tankeeksperiment, er at forestille sig at korpus *er* et billede af hele sproget, og kun se på de søgninger der faktisk giver et match, dvs. sortere støjen fra ved at udelade no-matches. Resultatet af det eksperiment fremgår af figur 4. Den overordnede sammenhæng er ikke grundlæggende forskellig fra den der fremgik af figur 3, men adskiller sig dog på to væsentlige punkter. Hvis man betragter X-aksen, kan man se at der skal betragtelig færre ordforekomster til for at dække søgningerne, netop omkring de 600.000 som er

antallet af forekomster i det undersøgte korpus. Det andet forhold man skal hæfte sig ved, er at det nu er muligt at nå helt op på de 100 %, hvilket naturligvis ikke bør komme som en overraskelse eftersom det kun er de søgninger der har et match i korpus, der er taget i betragtning.



Figur 4: Sammenhæng mellem korpusfrekvens og korpusdækning uden no-matches

Det vigtigste at hæfte sig ved i begge figurer er dog selve kurvens form. Man kan se at der skal overordentlig mange ordformer til for at komme fra 80 % til 100 %. Desuden noterer man sig at den absolutte frekvens for ordformerne bliver meget lav når man passerer ca. 100.000. Kurven knækker og er derefter meget længe om at nå det sidste stykke op til de 100 %. Det udlægger vi på den måde at værdien af at bruge korpus som princip i lemmaselektionen aftager efter det punkt hvor kurven begynder at flade ud. Når man befinder sig i den øverste del af kurven, vil det næste ord i rækkefølgen blot have 1 eller 2 forekomster i det korpus vi har anvendt til undersøgelsen, og det vil i praksis sige at det ene ord kan være lige så godt som det andet. Da det netop er i det område vi befinder os med DDO, er det en vigtig viden for os.

Den første delkonklusion vi kan udlede af undersøgelsen, er derfor at korpus kan være et godt redskab til lemmaselektion, men ikke er lige godt i alle faser af et projekt. Især er det vigtigt til de første ca. 100.000 ordformer, hvorefter værdien er aftagende, og efter ca. 200.000 ordformer har de næste mange former bare 1 eller 2 forekomster, og korpus er da ikke specielt velegnet til at afgøre hvad der udvælges. I den sammenhæng er det selvfølgelig vigtigt at skelne mellem ordformer og opslagsord. For nogle opslagsord, de ubøjelige, er der sammenfald mellem ordform og opslagsord, mens andre har ganske mange ordformer, bøjninger og stavelsesvarianter, pr. opslagsord. I gennemsnit er der ca. 3,5 ordformer pr. opslagsord i DDO.

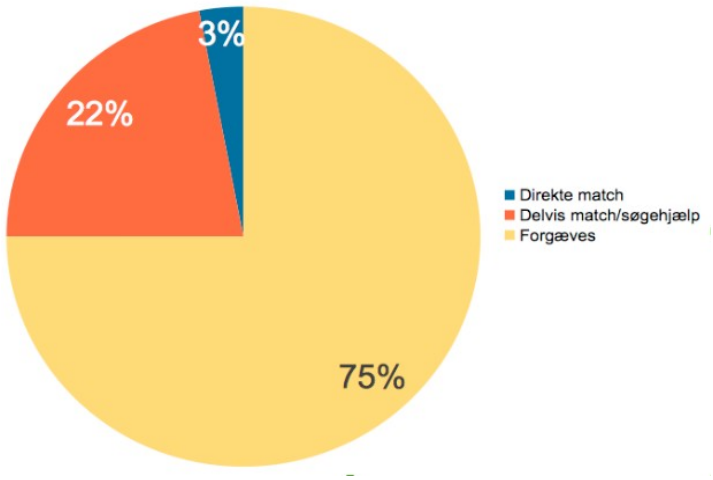
3. Indholdet af søgeloggen

Lad os derfor gå over til at se på den anden mulighed, nemlig hvad det er brugerne faktisk slår op. Ikke mindst er det af interesse at afdække hvad der søges forgæves. Forgæves søgninger defineres her som søgninger der ikke har et direkte match med en post i ordbogsbasen. Det kan synes ligetil, men der kan alligevel være grund til at skelne mellem forskellige typer af manglende match:

1. søgninger der ikke har et direkte match, men hvor brugeren alligevel får et resultat. Det er tilfældet hvis man søger vha. jokertegn, eller hvis man skriver flere ord i søgefeltet.
2. der er intet match og vises heller ikke noget resultat, men brugeren får alligevel hjælp til at komme videre, enten med alternative forslag fra "mente du"-funktionen eller med en besked om at ordet findes i en anden ordbog, i dette tilfælde *Ordbog over det danske Sprog*.
3. søgninger efter nonsens-ord eller ord der er så fejlstavede at "mente du"-funktionen ikke har forslag til hvad der kan være ment.
4. søgninger efter ord som ikke er i ordbogen i forvejen.

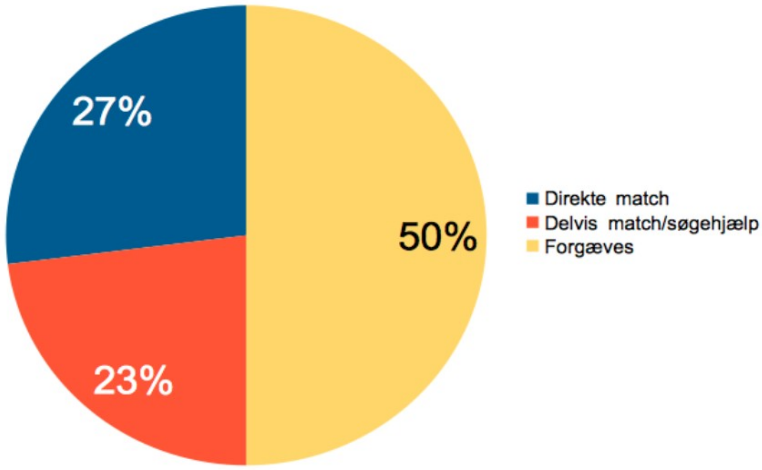
Den første type vil næppe opleves som forgæves af brugerne. Selvom strengen teknisk set ikke har et eksakt match i basen og derfor er omfattet af definitionen, vil brugerne alligevel ofte få svar på det de ledte efter. Den anden type vil heller ikke altid opleves forgæves, for brugerne får hjælp i form af forslag til hvor de kan finde det ønskede. De to sidste typer er derimod reelt forgæves, både for brugeroplevelsen og i teknisk forstand. Her gælder det om at skelne støjen fra de rigtige ord som er potentielle lemmakandidater.

For at få et indtryk af hvordan de forgæves søgninger fordeler sig på de 4 typer, udvalgte vi 100 tilfældige no-matches og undersøgte fordelingen. Resultatet vises i figur 5. I figuren kaldes type 1) for direkte match fordi brugeren får vist et resultat af sin søgning. Derimod er 3) og 4) slået sammen fordi det ikke er nemt at afgøre hvornår en søgestreng skal regnes som nonsens eller fejl, og hvornår som et rigtigt ord. Mere om dette senere.



Figur 5: Fordelingen af no-matches på typer blandt 100 tilfældige

Det fremgår af figur 5 at omkring en fjerdedel er de halve eller hele succeser, mens tre fjerdedele udgøres af reelle no-matches. I stedet for at se på et tilfældigt udvalg kunne man også se på de hyppigste forgæves søgninger. Det har vi gjort i figur 6.



Figur 6: Fordelingen af no-matches på typer blandt de 100 hyppigste

Blandt de 100 hyppigste forgæves søgninger giver halvdelen alligevel enten et rigtigt resultat eller hjælp til brugeren, mens der nu kun er 50 % forgæves søgninger. Forskellen mellem de hyppigste og de tilfældige forgæves søgninger skyldes mest at der blandt de hyppige er flere ikke-forgæves søgninger med jokertegn (*z*, *z, c*, x**), samt flere søgninger efter faste udtryk (*i hvert fald, i dag, til stede, vi får se*), efter former der tidligere har været officielle (*linie, krem, bolche, elvte*), er talesprogsnære (*osse, pive*) eller mangler forkortelsespunktummer m.m. (*su, email, ac, pt*). Disse søgninger fører alle til et resultat gennem vores søgealgoritmer selvom de ikke har et direkte match med en form i basen. Omvendt udgør den store mængde af ord der kun er slået op en enkelt gang, forholdsvis meget mere på listen over tilfældige søgninger, jf. nedenfor.

Hvis vi dernæst ser på hvad de forgæves søgninger dækker over, og først ser på de 100 hyppigste, er der blandt de 50 reelt forgæves søgninger hele 8 proprier, herunder både personnavne, geografiske navne og firmanavne som fx *Jørgen, Danmark* og *TDC*, mens de øvrige no-matches fordeler sig mellem nonsenssøgninger og mulige opslagsord. Det kan være svært at afgøre hvor grænsen mellem nonsens og oprigtige søgninger skal sættes, og derfor er der ikke skelnet mellem dem i figur 6. Nogle af de søgninger der ikke giver mening umiddelbart, kunne fx godt stamme fra deltagere i forskellige sprogspil (*Wordfeud, Wordeye, Ordkamp* el.lign.) der tjekker om der findes et ord *cu, po, zi* eller *xa* eller hvilke irriterende bogstaver de nu måtte sidde med. Det kan også være opslag der er foretaget automatisk, måske fra hackere, det kan være svært at vide. Pointen er at man er nødt til at have mere viden om søgeintentionen for at kunne klassificere en streng helt korrekt, og vores opdeling kan derfor ikke blive mere end et skøn, med de muligheder for fejl som det indebærer. Vi har skønnet at ca. 35 af søgningerne er nonsenssøgninger (eksempler er *bibhld, nyopmb, cu, xa, config, ce, tilfr, tr, tac, npop, nbceq1visse, za, ci, cu , zo, ma, xe*), mens højst 6 kan opfattes som reelle lemmakandidater: *værv, %, tf, lol, malacostraca, tac*. De færreste vil nok mene at det er umistelige ord: *%* kan også være et jokertegn, og *værv* kan måske være en fejlstavning for *hverv.* Heller ikke det kan man vide med sikkerhed når kun selve den indtastede søgestreng er til rådighed. Hvad brugerne har ment da de skrev strengen, får vi aldrig at vide. Alt i alt er det fortrøstningsfuldt for redaktionen at der ikke optræder mange helt oplagte lemmakandidater blandt de hyppige no-matches som den kan kritiseres for ikke at have taget med.

Ser vi også på de 100 tilfældige no-matches, tegner der sig til dels det samme billede og til dels noget andet. Der optræder igen en del proprier og også en del nonsens-søgninger, men her er også betydelig flere potentielle lemmakandidater, fx *uvirkelighedsfølelser, brancheafdeling, faghøjskoledanmark, enzymgigantens, pcbank, forskningskronerne, sluddertante, lydforhør, udførelsesmetoder, evg, jubilæumsmiddag, lavadal, differentiatorer, hvilestue, vigilante*. De fleste af ordene er sammensætninger, nogle af dem lejlighedsdannelser, mange af dem er betydningsmæssigt gennemskuelige rækkedannelser, men der er også neologismer og fremmedord iblandt (*differentiatorer, vigilante*). De hører dog typisk til i den del af korpus hvor kurven flader ud i figurerne 3 og 4, dvs. med 0, 1 eller 2 forekomster i korpus, og udgør ikke centrale ord i sproget, men måske de ord som man efter korpusudvælgelsesmetoden ville vælge som de næste.

Endelig er der en restgruppe bestående af fejlstavninger (fx *fenimonal, illustraioner, tilrekkligt, forrørt*) og andre ord som kan være lidt svære at fortolke, men om hvilke det også gælder at de typisk kun har en enkelt korpusforekomst (fx *femtinul, kulinaristisk, chemotolgy, caldérons, turos*).

Selvom der altså er lidt flere reelle lemmakandidater blandt de tilfældige no-matches, er det

dog ret perifere ord der dukker op i søgeloggen. En forklaring på hvorfor der ikke er flere, kan være at der redaktionelt allerede er foretaget en del ændringer på baggrund af no match-søgningerne. I 2010 konstaterede vi at no match-listerne indeholdt en del bøjede former som ikke gav et match i basen. De blev derefter tilføjet, hvorfor de naturligvis ikke længere optræder på no match-listen. Det drejer sig primært om verber i præsens participium og substantiver i genitiv. Endvidere er basen blevet beriget med flere almindelige fejlstavninger end dem vi allerede havde. Disse er jo ikke synlige for brugeren, men gør at "mente-du"-funktionen oftere kommer med det rigtige alternativ. Det gælder ikke mindst sær- og sammenskrivningsfejl. Forbedring af "mente-du"-funktionen bør i øvrigt løbende raffineres til at klare andre slå- og stavefejl.

Også blandt de reelle lemmakandidater har vi i løbet af perioden optaget nogle af ordene, bl.a. neologismer som fx *swag* og *hipster*[2]. Blandt de resterende ord er der en del relativt hyppigt søgte fremmedord som godt kunne være kandidater til at blive optaget. De er ganske vist ikke velrepræsenterede i korpus, men det kan jo også skyldes at korpus ikke er stort nok, eller at det indeholder for få akademiske, tekniske eller faglige tekster hvori den slags ord optræder.

De forholdsvis mange proprier der optræder blandt de forgæves søgninger, bør også få redaktionen til at overveje sine principper fordomsfrit. Argumentet mod at have proprier i betydningsordbøger har traditionelt været pladsøkonomi og vanskelighederne med at afgrænse ordstoffet, men det argument vejer ikke tungt i den digitale verden. Det kan være svært for brugerne at forstå at de kan slå *parisisk, cubaner* og *keynesiansk* op, men ikke *Paris, Cuba* og *Keynes*, og retskrivningsmæssigt kan navnestoffet volde mindst lige så store problemer som resten af ordforrådet. Det gør i hvert fald indtryk at der er så mange proprier blandt de ord brugerne prøver at slå op.

Endelig er der de brugergenererede forslag, funktionen "Send et ord", hvor brugerne kan sende indberetninger til den database som administreres af Dansk Sprognævn og Det Danske Sprog- og Litteraturselskab i fællesskab. Ikke overraskende er der et vist overlap mellem no-matchene og brugerindberetningerne; eksempler fra basen er bl.a. *egalitær, webinar, spilleliste, stomp, zumba*, *app*, *skype (vb.), blingbling,* alle ord der siden er blevet oprettet som opslagsord og nu kan slås op i ordbogen. Værdien af de indsendte brugerforslag er vekslende, mange af dem er næppe oprigtigt mente, andre har ad hoc-karakter, men forslag som indsendes af flere brugere uafhængigt af hinanden, fortjener at blive taget alvorligt. Basen indeholder for tiden over 6.000 poster.

4. Konklusion

Gennemgangen har set på forskellige kilder til lemmaselektion. Analysen af ordformernes fordeling i korpus viste at korpus er udmærket som kilde, men værdien er størst ved de første ca. 20.000 opslagsord, hvorefter værdien gradvis aftager. Derefter er det en god idé også at inddrage andre kilder. Søgeloggen er én mulighed, hvor lister over brugernes forgæves søgninger kan være en kilde til gode kandidater, især hvis de kombineres med brugernes indsendte forslag. Uanset hvilken fremgangsmåde der vælges, bør det ske i kombination med en redaktørs professionelle skøn, især for den del af ordforrådet der ligger uden for almensprogets almindeligste ord. Vi har set at ordene i den mere perifere del af ordforrådet typisk har en lav korpusfrekvens, og ved at kombinere de forskellige metoder bliver det bedste princip for lemmaselektion hvad man kan kalde en slags computerstøttet introspektion. Netop fordi det ene ord er omtrent lige så godt som det andet, set fra en korpuslingvistisk synsvinkel, bliver redaktørens rolle så meget mere afgørende. Trods større automatisering og brug af it i ordbogsarbejdet er der i høj grad brug for leksikografer og deres kompetence.

Det er i sig selv interessant at der er overordentlig mange ord med en korpusfrekvens på 1 eller 2 der bliver slået op. Hvis man har som ambition at hjælpe brugeren med at finde alle de ord som de prøver at slå op i ordbogen, betyder det at lemmabestanden i DDO bør udvides markant. Korpus kan her bruges til at finde yderligere kandidater som for de flestes vedkommende bør beskrives.

Lars Trap-Jensen
Det Danske Sprog- og Litteraturselskab
ltj@dsl.dk

Litteratur

Bergenholtz, Henning & Mia Johnsen (2005): Log Files as a Tool for Improving Internet Dictionaries. I: *Hermes* 34, s. 117-141.

Bergenholtz, Henning & Bjarni Norddahl (2012): Ordbogsartikler, som ingen læser. I: *LexicoNordica* 19, s. 207-223.

DDO = *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab. Onlineversion: http://ordnet.dk/ddo

de Schryver, Gilles-Maurice & David Joffe (2004): On How Electronic Dictionaries are Really Used. I: Geoffrey Williams & Sandra Vessier (red.): *Proceedings of the Eleventh EURALEX*

International Congress, Euralex 2004. Lorient, France. July 6-10. Volume I. Lorient: Université de Bretagne, s. 187-196.

Lorentzen, Henrik & Sanni Nimb (2011): Fra krydderkage til running sushi – hvordan nye ord kommer ind i Den Danske Ordbog. I: Margrethe Heidemann Andersen & Jørgen Nørby Jensen (red.): Nye ord, Sprognævnets Konferenceserie 1, Dansk Sprognævn, s. 69-85.

Ordbog over det danske Sprog 1-28 (1918-56) med *Supplement* 1-5 (1992-2005), København: Det Danske Sprog- og Litteraturselskab og Gyldendal. Onlineversion: http://ordnet.dk/ods.

Norling-Christensen, Ole & Jørg Asmussen (1998): The Corpus of The Danish Dictionary. I: Lexikos. Afrilex Series 8, s. 223-242.

1 Denne artikel bygger på resultater af en undersøgelse gennemført i efteråret 2013. Undersøgelsen blev udført i samarbejde med Nicolai Hartvig Sørensen og Henrik Lorentzen, Det Danske Sprog- og Litteraturselskab, og præsenteret ved konferencen eLex 2013, electronic lexicography in the 21st century: thinking outside the paper (se http://eki.ee/elex2013).

2 Principielt er det et problem hvis der manuelt er tilføjet mange ord fra no match-listen til ordbogen i perioden fordi kvaliteten af de resterende lemmakandidater på listen derved underspilles. Problemets omfang bør dog heller ikke overdrives da det ikke drejer sig om noget stort antal, og desuden ville de nævnte eksempler også blive udvalgt på grundlag af deres frekvens i det moderne korpus.