NyS

Titel:	Manuel og maskinel excerpering af neologismer
Forfatter:	Jakob Halskov og Pia Jarvad
Kilde:	*NyS – Nydanske Sprogstudier* 38, 2010, s. 39-68
Udgivet af:	NyS i samarbejde med Dansk Sprognævn
URL:	www.nys.dk

NyS 38

Søgbarhed

Artiklerne i de ældre NyS-numre (NyS 1-36) 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.

Manuel og maskinel excerpering af neologismer

JAKOB HALSKOV & PIA JARVAD

INDLEDNING

I denne artikel vil vi beskrive arbejdet med Dansk Sprognævns Ordtrawler som er udviklet af Jakob Halskov. Ordtrawleren har til formål automatisk at finde nyordskandidater (konstruktioner som kandiderer til at blive regnet for nye) i de meget store tekstkorpusser som nu til dags er tilgængelige på nettet, og dermed bidrage til at følge sprogets udvikling. Vi har foretaget en formel og kvalitativ evaluering af dette automatiske excerperingssystem, og vi diskuterer styrker og svagheder ved automatiseret kontra menneskelig excerpering. I evalueringen udgør menneskelige vurderinger den guldstandard som systemet holdes op imod.

Hvad er et nyt ord?

Klatfravær, *mælkeskandale*, *boliggaranti ver. 2.0*, *bøllebank*, *udsatteråd*, *caucusvalg*, *popsprog*, *bandekrig*, *bitterfisse*, *burkaudvalg*, *dialogmedicinering*, *feelgood*,*grønvask*, *kant* (ny brug), *klimacertifikat*, *kontraktpolitik*, *sundhedskort* (= sygesikringskort), *nederen*, *oldschool*, *prettyboy*, *tudefjæs*, *uhellig alliance*, *underfrankeret* (billedlig brug om svagtbegavet), *überseksuel*

En tilfældig række af nye ord fra de seneste år. Nogle af dem er benævnelser for nye ting og fænomener, fx *udsatteråd* 'kommunalt råd for varetagelse af udsattes (fx hjemløse) tarv', *grønvask* 'det at varer o.l. fremstår som mere økologiske og klimavenlige end de i realiteten er', *mælkeskandale* (om skandale i Kina hvor der blev tilsat melamin i mælken med skrækkelige sundhedsmæssige konsekvenser). Andre er nye ord for et fænomen som kendes i forvejen og som i forvejen kan have en benævnelse for sig, fx *klatfravær* og *sundhedskort* (= *sygesikringskort*). Gamle ord får tillagt nye betydninger, fx *kant* og *underfrankeret*. Til nye ord regnes også at ordet bruges på en ny måde, fx *dumpe eksamen*

hvor det traditionelle er *dumpe til eksamen*. Udtryk som *boliggaranti ver. 2.0*, *uhellig alliance* har, selvom de består af to ord, som helhed en fast betydning og fast form, og de er således nye faste udtryk, og de hører med i beskrivelsen af nye ord.

Nye ord og ny brug af gamle ord er nyheder i forhold til det inventar af ord som findes i forvejen, dvs. at vi regner nye ord for nye når de ikke står i gængse ordbøger (fx Den Danske Ordbog, Retskrivningsordbogen, Nudansk Ordbog, og ikke mindst Nye ord i dansk på nettet fra 1955 til i dag[1] og Dansk Sprognævns Samling). I excerperingen undgås de ord som ikke antages at ville blive etablerede i ordforrådet. Det er banale sammensætninger, lejlighedsdannelser og herunder kometord (se nedenfor).

Banale sammensætninger er fx *klimakonference*, *klimakunst*, *klageantal*, *værdipapirsammensætning*, *risikoappetit*. De er dannet af sprogets byggeklodser efter de regler som vi opbygger et ord på, og de er gennemskuelige og uproblematiske at forstå hvis man kender førsteled og andetled. De er karakteriseret ved at de kan dannes nu, men de kunne lige så godt have været dannet for 50 år siden hvis der havde været behov for det.

Lejlighedsdannelser (øjebliksdannelser eller individualdannelser som de også kaldes) har et mere tilfældigt præg, fx *tefest* (jf. *vinfest*), *tesøster* (jf. *kaffesøster*), tetår (jf. *kaffetår*), og *tetelt* (jf. *øltelt*). Sådan set er *tefest* mfl. fuldgode regelret dannede danske ord - men de er dannet til denne lejlighed her og nu og bliver næppe brugt nogensinde mere. Disse ord forstås umiddelbart, men det bagvedliggende ord som fx *kaffesøster*, *øltelt* med disse ords bibetydninger gør forståeligheden større. En *tesøster* skal derfor ikke forstås som en ‚søster der holder af te', men derimod som ‚person som holder meget af te'.

I Politiken 4.11.2009 kunne man læse en politikers udtalelse om sit eget barns gåen i privatskole versus politikerens officielle udtalelser om folkeskolen:

> Når det gælder mit barns uddannelse, er min morkasket trods alt vigtigere end min politikerkasket.

Her er der forudsat en del viden om betydningen af kasket (jf. Nye ord på nettet: "kasket *sb.* (1977) (billedlig brug i forbindelser som have to el. flere kasketter på, skifte kasket) som udtryk for en funktion, et hverv, at have en dobbeltrolle)".

En variant af lejlighedsdannelserne er ord som bliver almindelige i sprogsamfundet og forsvinder igen inden for en kort periode. De kaldes kometord. Det er fx *burkaudvalg* og *klimakaravane* 'bus som kører rundt i landet i 2008-9 med oplysning om den globale opvarmnings betydning for klimaet', og *mælkeskandale* (se mere herom nedenfor). Sådanne kometord har stærk affinitet til bestemte hændelser og kan ikke forstås uden at man kender til hændelserne. De er leksikaliserede. De har potentiale til at blive varige tilskud, men det finder man først ud af når der er gået et stykke tid.

Prægnante ord er de nye ord som bliver varige i sproget, mens de ovennævnte er uprægnante og uinteressante i nyordsperspektivet (men ikke i et orddannelsesperspektiv). De uprægnante er der rigtig mange af; det kommer vi ind på senere i artiklen. Det er selvsagt de prægnante nye ord og udtryk som har størst interesse for beskrivelsen af sprogets udvikling.

EXCERPERING OG KILDER

Dansk Sprognævn blev oprettet i 1955 og fik som én blandt flere opgaver at følge med i ordforrådets udvikling. Der står således i bekendtgørelsen om Sprognævnet at en af nævnets arbejdsopgaver er: „At følge det danske sprogs udvikling navnlig ved at indsamle og registrere nye ord, ordforbindelser og ordanvendelser, herunder forkortelser"[2]. Dansk Sprognævn er den eneste institution herhjemme der systematisk arbejder med indsamlingen af nye ord, og Sprognævnets ordkartotek, både i dets elektroniske og fysiske form bruges som udgangspunkt for opdatering af mange ordbøger, leksikoner og andre opslagsværker. I Sprognævnets ordkartotek findes også andet end nye ord. Der excerperes (udtrækkes) varianter af de officielle staveformer og bøjningsformer af hensyn til arbejdet med retskrivningsordbogen, og der excerperes sprogbrug og syntaks af hensyn til den sproglige

rådgivning, men her er fokus de nye ord.

For at indsamle nye ord og ordanvendelser læses der aviser, ugeblade, tidsskrifter, bøger og mange andre typer tekster som repræsenterer forskellige genrer. Teksterne er fra hele landet og er fordelt på emner og efter hvilken aldersgruppe de forskellige tekster retter sig imod. Der læses lige fra romaner og digtsamlinger til varekataloger, fra dagbladet Politiken til Folkebladet, Dagblad for Vejen, Brørup, Holsted og omegn, fra tidsskriftet Press til Statstidende og fra magasinet Vi unge til sundhedsbladet Helse sammen med dameblade, mandeblade og etiketter på madvarer. Der lyttes til Danmarks Radio, lokalradioer, de forskellige tv-stationer, og til hvad der siges i de daglige samtaler. Den sidsttilkomne kilde til nye ord er de nye elektronisk bårne medier, chat, blogs mv. på nettet, twitter, konsolspil osv. Teksterne har gennem årene fordelt sig således:

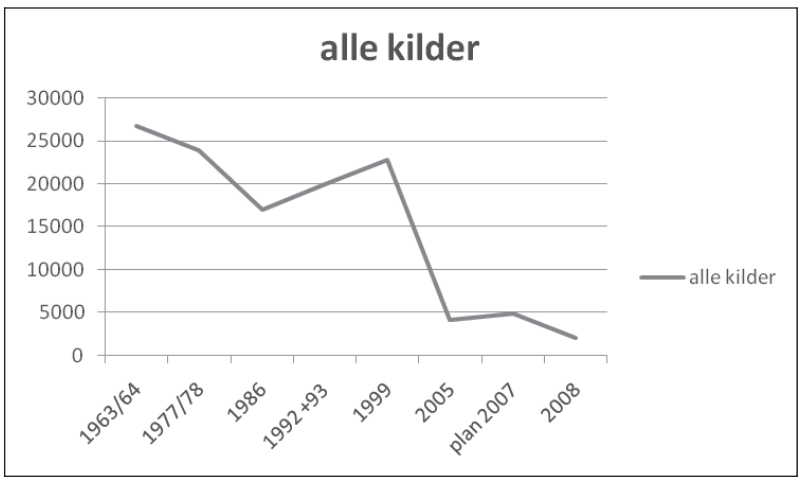
TABEL 1. FORDELING PÅ KILDER I PROCENT.

	1963/64	1977/78	1986[3]	1992+93	1999	2005	Plan 2007[4]	2008
avis, kbh	52	61	24	49[5]	49	60	10	68
avis, provins	3	17[6]	12		13	4		6
distriktsblade[7]	4	2	10	0[8]	2	8		2
kortgenre[9]	2	1	10	3	3	1	18	2
ugeblade			6[10]	17	4	11	13	13
faglige ugeblade[11]	4		8		3	1	18	4
officielt sprog[12]	9	5	5	5	1	1	8	0
bøger[13]	12	4	6	9	12	8	16	0
fremmedsprog[14]		11	6	8	1	1		0
tidsskriftet "Bogmarkedet"			8		4	0		0
mundtlig	1		5	9	4	3	8	0
andet	11		1	1	3	1		4
internet						0		1
metasprog							8	
	98	101	101	101	99	99	99	100

Det drejer sig om kilder excerperet – og opgjort årligt – i mere end 50 år. Der er derfor fluktuationer i måden at opgøre kilderne på, og der er i denne opgørelse slået kilder sammen af hensyn til sammenligneligheden. Tallene i tabel 1 er i procent af det samlede antal excerpter det pågældende år. Alle procenter er rundet op til nærmeste hele tal, og derfor kan tallet i alt afvige fra 100 %. Der er mange fodnoter som redegør for baggrunden for tallet/kilden/resultatet.

Kilderne til excerperingen er begrænsede og udvalgt af excerpisterne – nogle gange efter erfaring med hvor man finder mest, fx at aviser rummer flest nye ord. Kilderne som har været excerperet i Dansk Sprognævn siden begyndelsen af 1960'erne, fordeler sig som man kan se det i tabel 1. Som man kan se, er aviserne langt den største kilde. Det skyldes at aviserne har nyhedsformidling som primær funktion, både nyheder i politik, samfundsforhold og kultur - og som følge deraf er det der de nye ord oftest dukker op på skrift første gang. Det er også i almindelighed aviserne der formidler de nye ord videre til andre medier. Men det er ikke nok at notere sig ordet den første gang det optræder. Ordet skal verificeres som nyt, som brugbart i mange genrer (alle genrer?) og blandt mange forskellige grupperinger af mennesker både socialt og geografisk, for at man kan sige at ordet tilhører det almene, fælles sprog i Danmark.

FIGUR 1. ANTAL EXCERPTER FORDELT PÅ ÅR FRA ALLE KILDER



I de første år var excerperingen beskeden, og finansåret 1962/63 var det første hvor man kom op på de magiske 20.000 excerpter som man i mange år anså som det optimale for at excerperingen var dækkende. Dette antal holdt sig helt frem til år-2000-skiftet hvor de første store maskinlæsbare tekster blev tilgængelige, og hvor et nyt ord kunne verificeres med hensyn til udbredelse, betydning mv., som den manuelle excerpering tidligere havde haft som opgave. Den manuelle excerpering har derfor kunnet koncentrere sig om udelukkende at finde nye ord. Men faldet er så stort at man næppe kan sige at der excerperes tilstrækkeligt. Faldet har i høj grad at gøre med at de to storexcerpister (Jørgen Eriksen og Arne Hamburger) blev pensionerede og derefter nedtrappede deres pensionistarbejde – og dette afløstes ikke af øget excerpering internt i Sprognævnet. I plan 2007 (se tabel 1 med fodnote) opstilles et minimumskrav til excerperingen, men desværre har prioritering af arbejdsopgaver i Sprognævnet i øvrigt gjort at planen ikke har været mulig at gennemføre. Derudover er 2008-resultatet alene udtryk for det som kom i den elektroniske samling, ikke det som rent faktisk er indsamlet. I 2009 er der således store forhåbninger til Jakob Halskovs Ordtrawler og dens bidrag til øgning af nye ord i samlingen. I 2008 iværksattes også en funktion på nettet hvor det var muligt for almindelige mennesker at indberette ord som opfattes som nye. Det har dog ikke givet synderligt resultat.

NYHEDSMARKERINGER

En almindelig udbredelsesmåde for et nyt ord er at det dukker op i en avis, og tingen eller fænomenet bliver omtalt, det bliver forklaret og måske sat i citationstegn eller løftede kommaer, og brug af kursiv eller anden grafisk særmarkering ses, fx

> Italien sender nu 'grundløse asylansøgere' tilbage. (Politiken 25.1.2009)

Ofte er nye ord forklaret rent betydningsmæssigt, og skribenten kommenterer ordet. Ikke sjældent sættes et "såkaldt" foran ordet, fx

> Meget af denne vold er såkaldt opdragelsesvold. (Politiken 2.11.2009)

Hvis det er et ord fra fremmed sprog, kan det være forsøgt oversat, fx

> I dag er det anonyme *ghostwriters*, der lægger ord i munden på toppolitikere ... nutidens spøgelsesskribenter. (Politiken 2.11.2009)

Sammensætninger er ikke sjældent skrevet i to ord eller med bindestreg mellem leddene. Alle disse træk er signaler for at det pågældende ord er en nyhed.

Er det nye ord, der betegner fænomenet eller tingen, et ord der er blivende, så anvendes det i en periode med disse typer af nyhedsmarkering, senere forsvinder nyhedsmarkeringen, og ordet indgår på lige fod med det øvrige ordforråd. Ordet bliver så også almindeligt i kilder som ikke er nyhedsformidlende, fx ugeblade og bøger for senere evt. at blive brugt i officielt sprog som love og bekendtgørelser. Fx blev ordet *knallert* brugt om det man officielt kaldte *cykel med hjælpemotor* allerede omkring 1950, men først med 1976-færdselsloven blev ordet brugt i officiel sammenhæng.

MANUEL KONTRA AUTOMATISERET EXCERPERING

Nyhedsmarkeringer er gode clues for den nye excerpist, men nyhedsmarkeringer er der ikke altid. Eftersom det nye ord ikke findes i forvejen, er der ikke egentlige metoder man kan benytte når teksten læses for at finde de nye ord. Derimod bruger excerpisten sin viden om modersmålets orddannelse og ordforråd, og jo større erfaring med arbejdet med nye ord, jo bedre en excerpist. Ord som er i kikkerten, undersøges i ordbøger, ordsamlinger mv. for at verificere at ordet ikke er gammelt. Nudansk Ordbogs seneste udgave har været rettesnor – hvis ordet ikke var dér, blev det excerperet.

Den manuelle excerpering repræsenterer en høj grad af abstraktion i forhold til et ubearbejdet korpus, og resultatet, excerpterne på seddel eller i database, kan i mange tilfælde betragtes som midtvejs mellem korpus og ordbog. I den manuelle excerpering har excerpisten som modersmålstalende viden om sproget og kan derfor se bort fra grafiske elementer hvis det er nødvendigt, og hun kan lemmatisere (dvs. henføre til opslagsform) og normalisere (dvs. slå ortografiske varianter

sammen), og hun kan vurdere ordet og dets betydning i forhold til konteksten samtidig med excerperingen, og hun har tillige viden om omverden. I den manuelle excerpering filtreres således en stor del fra helt umiddelbart.

Atkins & Rundell (2008: 51) hævder, med visse forbehold, at det er en smal sag at excerpere nye ord automatisk, men i Halskov og Jarvad (2009, 2010) redegøres der for de ganske store problemer der er med at fremfinde egnede kandidater; der er simpelthen for meget støj ved at Ordtrawleren finder for mange kandidater som ved et nærmere eftersyn ikke ville blive fundet excerperingsværdige i en manuel excerpering. Ligeledes er antallet af fundne "rigtige" kandidater for lille når søgningen foregår med visse filtre (filtre der fx udelukker kandidater som ikke er nyhedsmarkeret med citationstegn eller *såkaldt(e)*).

HVORDAN EXCERPERER EN MASKINE?

Der findes et hav af natursprogsbehandlingssystemer til automatisk fremfinding af termkandidater (altså potentielt fagsproglige udtryk) i et tekstkorpus, men der er nærmest ikke publiceret nogen tekniske detaljer om systemer som kan excerpere almensproglige (eller fagsproglige) *nydannelser*. Dette betyder imidlertid ikke at sådanne systemer ikke eksisterer. Det engelske APRIL-projekt (*A knowledge-rich tool for the analysis and prediction of innovation in the lexicon*)[15] er måske det mest velkendte af slagsen, men forfatterne er også bekendt med et lignende projekt som forskere ved Universitetet i Bergen står bag. Desværre er det svært at vide præcist hvilke teknikker der benyttes, og dermed afgøre i hvilket omfang det nærværende system (Ordtrawleren) udgør en decideret nyudvikling. Ordtrawleren består i sin nuværende form af en håndfuld tekstbearbejdningsprocedurer (små programstumper), en stor database (til lagring af tekstmateriale, filtre og nyordskandidater), en korpusservice og en simpel brugergrænseflade til forskellige korpusværktøjer.

Før Ordtrawleren kan excerpere, underkastes de elektroniske tekster en række automatiserede behandlinger i en bestemt rækkefølge.

1. *Tokenisering*: Brødteksten deles op i sætninger, og sætningerne hakkes op i en sekvens af ordformer.
2. *Part of Speech-tagging*: Hver ordform tildeles automatisk en ordklasse (her anvendes de forenklede Parole tags (Keson 1998).
3. *Lemmatisering*: Hver ordform tildeles automatisk et lemma (her anvendes en udfoldet version af Retskrivningsordbogen 2001).
4. *Indeksering*: De enkelte oplysninger om hver ordform, dvs. formen, ordklassen og lemmaet lagres og indekseres i en database så man hurtigt kan gennemsøge store mængder tekst for bestemte mønstre (her anvendes *Corpus Workbench*[16]-formatet).
5. *Filtrering/sortering*: Inventaret af samtlige forskellige ordformer (det samlede ordforråd) filtreres og/eller sorteres ved hjælp af ord- og frekvenslister over allerede kendte ord.

Den automatiske tokenisering, tagging og lemmatisering er naturligvis ikke fejlfri. Taggeren som anvendes, er beskrevet i Hansen (2000) hvor den vurderes at have en træfrate på 96,5 %, men "dog bliver kun ca. 80 % af alle ukendte ord gættet" (Hansen 2000: 7). Ord som ikke er indeholdt i værktøjets ordbog, volder altså særligt store problemer, og det er jo netop en delmængde af disse ord vi er ude efter. Problemet gælder i særlig grad den lemmatiseringsteknik som i øjeblikket anvendes af Ordtrawleren (punkt 3 ovenfor). Der anvendes nemlig en udfoldet version af Retskrivningsordbogen 2001, og ordformer som ikke kan henføres til et opslagsord i dette værk, lemmatiseres dermed ikke. Vi vender tilbage til denne problematik senere i artiklen.

Tabel 2 nedenfor indeholder et eksempel på hvordan sætningen, "At forbyde salg af tobak er ikke en måde at forlænge danskernes levetid med.", ser ud efter ovenstående automatiserede behandlinger.

TABEL 2. EN BEARBEJDET SÆTNING I DET SÆRLIGE CORPUS WORKBENCH-FORMAT

Ordform	***Ordklasse***	***Lemma***
At	UKONJ	at
Forbyde	V_INF	forbyde
Salg	N	salg
Af	PRAEP	af
Tobak	N	tobak
Er	V_PRES	være
Ikke	ADV	ikke
En	PRON_UBST	en
Måde	N	måde
At	UKONJ	at
Forlænge	V_INF	forlænge
Danskernes	N_GEN	dansker
Levetid	N	levetid
Med	PRAEP	med
.	TEGN	.

Når al tekstmaterialet (analysekorpusset) foreligger i ovenstående format, trækkes hele dets ordforråd ud (dvs. alle ordformer, også kaldet "types") og sammenlignes form for form med det ordforråd systemet kender i forvejen. Senere i artiklen vil vi beskrive hvilke eksisterende ordbøger og referenceværker der anvendes til filtrering og sortering.

Da den kunstige intelligens lader vente på sig, så har maskiner stadig vanskeligt ved at abstrahere med mindre de er blevet eksplicit programmeret til det. En maskine vil således som udgangspunkt opfatte stavevarianter som vidt forskellige ord og altså foreslå *U.S.A.* som nyordskandidat selvom den har formen *USA* i sin liste over allerede kendte ord. På tilsvarende vis skal en maskine også have detaljerede instrukser om hvordan den skal håndtere bindestreger, små/store bogstaver, citationstegn osv. I modsætning til menneskelige excerpister, for hvem det er helt naturligt at inddrage et ords kontekst, så kræver det temmelig avancerede programmeringsteknikker at få maskiner til at tage hensyn til den sproglige kontekst hvori et ord optræder. Således vil

ny brug af eksisterende udtryk, ny valens, nye flerordsforbindelser osv. være vanskelige at få en maskine til at identificere.

Der er altså ingen tvivl om at automatisk natursprogsbehandling er en vanskelig opgave, og i næste afsnit vil vi kort beskrive nogle fundamentale lovmæssigheder som nærmere kan forklare hvorfor det er særligt vanskeligt for en maskine at identificere sproglige nydannelser. Det drejer sig om lovmæssigheder som stammer fra to beslægtede grene af lingvistikken, nemlig datalingvistikken og især korpuslingvistikken.

DET KORPUSLINGVISTISKE PARADIGMESKIFT

Den korpuslingvistiske metode er en induktiv tilgang til lingvistikken i modsætning til mere deduktive tilgange som repræsenteret ved Noam Chomskys generative lingvistik og universelle grammatik. Den grundlæggende forskel på de to tilgange er at korpuslingvistikkens primære forskningsobjekt er den konkrete sprogbrug, altså *parole*, mens den generative tradition fokuserer på det abstrakte system af sprogbrugsregler der kan afledes af *parole*, dvs. *langue*. De to aspekter af natursprog hænger naturligvis sammen i et dialektisk kredsløb, men det korpuslingvistiske paradigmeskift som især tog fart i løbet af 1980'erne med britiske leksikografiprojekter som John Sinclairs COBUILD, gjorde det pludseligt lødigt i datalingvistiske kredse at tage udgangspunkt i sprogbrugen, hvilket det ikke havde været siden computerens absolutte barndom.

Inden for sprogteknologien (anvendt datalingvistik) og natursprogsbehandlingen (*Natural Language Processing*) betød paradigmeskiftet at eksempelvis maskinoversættelsessystemer i højere grad begyndte at anvende automatisk genereret sprogbrugsstatistik end manuelt kodede regler og at ontologiopbygning inden for Kunstig Intelligens også gradvist ophørte med at foregå manuelt, men i højere grad blev afledt af store mængder sprogbrug. En målbar effekt af paradigmeskiftet var at værktøjerne blev mere robuste og resurserne fik en langt større dækningsgrad.

Efterhånden som mere og mere tekst digitaliseres og tilgængeliggøres på internettet, så er korpussernes størrelser vokset fra ca. 1 mio. løbende ord i 1960'erne (jf. Brown-korpusset) til 100 mio. ord i 1990'erne (jf. *British National Corpus*[17] og det danske Korpus 90[18]). Efter årtusindskiftet er man nu begyndt at anvende delmængder af hele internettet som en slags brug-og-smid-væk-korpusser (Kilgarriff 2003) og empirisk funderede fraseologer kan dermed boltre sig i ngram-data[19] fra Google baseret på tekstmateriale fra ikke mindre end 1 billion hjemmesider[20].

Datarigeligheden kombineret med hurtigere computere har desuden medført at korpuslingvistiske studier ikke længere blot er korpus*baserede*, men i stigende grad tilmed er korpus*drevne* (Tognini-Bonelli 2001). Hvor det før var et spørgsmål om at efterprøve introspektivt funderede hypoteser empirisk, så tages der nu ofte direkte afsæt i automatiske analyser af store mængder data, og det er dermed computeren der præsenterer korpuslingvisten for en empirisk funderet hypotese som denne så introspektivt må tage stilling til. Et eksempel på en korpusdrevet sprogteknologisk applikation er Adam Kilgarriffs *Sketch Engine* (Kilgarriff et al. 2004) som automatisk kan danne en skitseagtig ordbogsartikel på basis af et lemma og et stort tekstkorpus. Et andet eksempel er Dansk Sprognævns Ordtrawler som evalueres i denne artikel.

Korpuslingvistikkens succes har altså i høj grad været betinget af den digitale revolution som informationssamfundet har medført. Det kræver imidlertid store mængder data at beskrive sprogbrugsmønstre for indholdsord. Og så snart man inddrager fx registervarians i sin analyse (jf. Biber 1998) eller bevæger sig over ordniveau for at analysere kollokationer, fraseologi og semantik, så øges behovet for empiri i endnu højere grad takket være ngram-analysens kombinatoriske eksplosion. Hvis der fx er 500.000 forskellige ordformer i et sprogs ordforråd, så kan der dannes 500.000 mulige unigrammer[21], men 500.000^2, dvs. 250 mia., (teoretisk set) mulige bigrammer[22] (fx "ispind drøm") og så fremdeles. Selvom det naturligvis kun er en brøkdel af bigrammerne som rent faktisk vil være grammatisk mulige (fx "gul drøm") og kun en brøkdel af disse som i realiteten vil forekomme med nogen nævneværdig frekvens i et korpus (fx "interessant drøm"), så illustrerer eksemplet at det kræver store mængder data

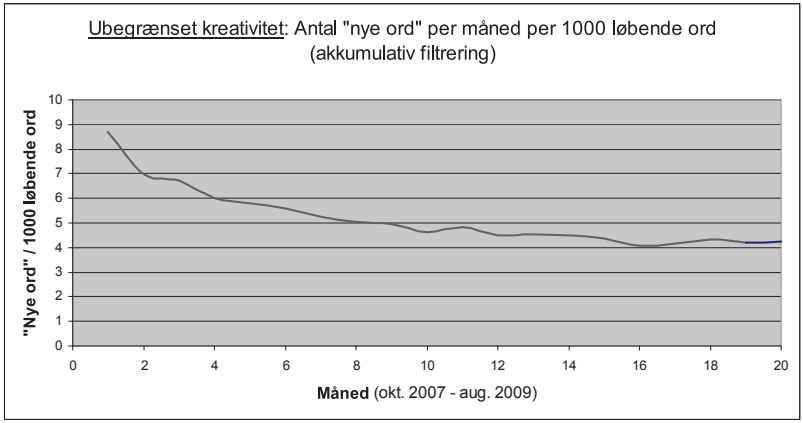
at kunne udtale sig om hvorvidt en nulforekomst af et givet ngram i et givet korpus betyder at konstruktionen er ugrammatisk eller blot mindre hyppig. Denne udfordring kaldes *the data sparseness issue* eller datautilstrækkelighedsproblemet og fører os videre til en kort bemærkning om Zipfs lov.

George K. Zipf beskrev allerede i 1935 (Zipf 1935) det fænomen at halvdelen af alle *types* (typisk indholdsord) i et givet korpus kun forekommer én gang, mens ganske få, særdeles hyppige *types* (typisk funktionsord) udgør ca. halvdelen af de løbende ord i ethvert korpus. Denne lovmæssighed betyder som sagt at det kræver store mængder data at kunne drage lødige konklusioner om sproglige fænomener som vedrører det åbne ordforråd, kollokationer og især nydannelser. Da nye ord begynder deres tilværelse med frekvensen 1, er det altså altafgørende at statistiske excerperingsteknikker ikke undervurderer betydningen af sjældne begivenheder i korpusset. Det bør nævnes at Zipfs lov primært gælder almensprog, idet de såkaldte subsprog, fx fagsprog, typisk har et fattigere og i princippet endeligt ordforråd (jf. begrebet *lexical closure* i McEnery (1996: 155-158)).

ORDTRAWLERENS EMPIRI: UBEGRÆNSET SPROGLIG KREATIVITET

I dette afsnit vil vi kort illustrere den grænseløse sproglige kreativitet og produktivitet som Ordtrawleren må forsøge at navigere i. Figur 2 nedenfor illustrerer hvordan Ordtrawleren måned for måned registrerer tusinder af ukendte ordformer i nyhedsartiklerne fra Infomedia. Det gør den også selvom man fjerner ca. 1,2 millioner allerede kendte ordformer fra et antal ordbøger og referencekorpusser og ser bort fra ikke-ord (der er tekststrenge med grafisk tegn som ikke er bogstaver; typisk URL'er og e-mail-adresser) og proprier og tilmed anvender akkumulativ filtrering af alle hidtil observerede ordformer.

FIGUR 2. UBEGRÆNSET SPROGLIG KREATIVITET.



Selv efter 20 måneder observerer systemet stadig mellem fire og fem ukendte ordformer per 1000 løbende ord. En enkelt måneds data fra Infomedia (ca. 7 mio. løbende ord) bidrager således med ikke mindre end 30.000 "nye ord". Citationstegnene tilkendegiver at en menneskelig excerpist naturligvis aldrig vil betragte mere end en brøkdel af disse tekststrenge som sproglige nydannelser der kan indgå i en nyordsordbog, men for en maskine er det anderledes vanskeligt at skelne skidt fra kanel. Tallet kan virke overraskende, for der er meget sjældent tale om stave- eller slåfejl i redigeret nyhedstekst, men læseren kan blot tænke på hvor mange sammensatte ord sprogbrugeren kan danne på basis af et enkelt mønster som tal-tal-(sejr/nederlag) (fx *32-14-sejren*).

Anvendes de førnævnte nyhedsmarkeringer, så kan antallet af kandidater imidlertid reduceres fra ca. 30.000 per måned til ca. 150 per måned. Man risikerer dermed at udmærkede nydannelser ignoreres fordi de ikke ledsages af en nyhedsmarkering, så denne teknik kan altså ikke stå alene. Hvis det analyserede tekstmateriale imidlertid er omfattende nok, så vil de fleste nydannelser sandsynligvis kollokere med et nyordssignal før eller siden (og måske især inden nydannelsen for alvor etablerer sig).

TO EKSPERIMENTER

I de følgende to afsnit vil vi beskrive resultaterne af to eksperimenter hvor Ordtrawleren har forsøgt at fremfinde nyordskandidater i to forskellige tekstsamlinger. De to eksperimenter har til formål at evaluere tre forskellige succeskriterier fra feltet *Information Retrieval*, nemlig *recall*, *precision og F-score*.

Mens *recall* og *precision* er defineret som følger, så udtrykker *F-score* balancen mellem de to succeskriterier.

$$recall = \frac{Antal\_relevante\_dokumenter \cap Antal\_fremfundne\_dokumenter}{Antal\_relevante\_dokumenter}$$

$$precision = \frac{Antal\_relevante\_dokumenter \cap Antal\_fremfundne\_dokumenter}{Antal\_fremfundne\_dokumenter}$$

I *Information Retrieval* er informationsenhederne typisk repræsenteret ved dokumenter, men i denne sammenhæng kan "dokumenter" i ovenstående formler oversættes med sproglige nydannelser. Systemets recall (dvs. genkaldelsesrate) er således givet ved delmængden mellem mængden af fremfundne nyordskandidater og mængden af samtlige "relevante" nye ord i materialet sat i forhold til mængden af samtlige "relevante" nye ord i materialet.For at evaluere systemets *recall* (dvs. genkaldelsesrate) er det nødvendigt (manuelt) at etablere en såkaldt guldstandard (dvs. facitliste) som omfatter samtlige nydannelser i materialet, med andre ord alt hvad systemet burde fremfinde i teksterne hvis det var perfekt. Med "relevante" menes der således nydannelser som er en del af denne guldstandard. *Precision* bliver i dette eksperiment dermed et udtryk for hvor stor en andel af alle maskinens fundne kandidater der svarer til kandidater i guldstandarden, *recall* fortæller hvor mange af guldstandardens kandidater maskinen finder i materialet, og *F-score* er et vægtet gennemsnit af de to andre mål.

Når man skal evaluere et systems *recall*, så må man vide præcis hvor mange relevante informationsenheder analysekorpusset indeholder. Manuel opmærkning er imidlertid meget tidskrævende, og derfor er det første eksperiments analysekorpus begrænset og består af et mindre

antal avisartikler som i alt udgør ca. 75.000 løbende ord.

For at evaluere systemets *precision* (dvs. træfrate) er det imidlertid ikke nødvendigt at gennemgå hele analysekorpusset og opmærke samtlige nydannelser. Det er tilstrækkeligt at analysere den sorterede liste nyordskandidater systemet genererer. Empirien for det andet eksperiment er således et omfattende antal avisartikler som udgør knap 100 mio. løbende ord.

1. EKSPERIMENT: EVALUERING AF RECALL OG PRECISION PÅ MINDRE KORPUS

Udgangsteksten er 177 korte avisartikler (ca. 75.000 løbende ord fra Jyllands-Posten i 2008) som en praktikant[23] ved Dansk Sprognævn manuelt har excerperet. Hun har opmærket nyordskandidater og sikret sig at disse ikke i forvejen var registreret i Sprognævnets ordbase. Nyordskandidaterne er vurderet i forhold til excerperingens mål, i dette tilfælde en nyordsbog. En seniorexcerpist har gennemgået de samme tekster, og det samlede resultat af excerperingen giver 252 ord som er guldstandarden. Formålet med denne guldstandard er at vurdere hvor stor en del af samtlige nydannelser i avisartiklerne Ordtrawleren formår at identificere. Af de 252 nyordskandidater er 33 nye ordforbindelser, fx *gå kort*, og 11 er ny betydning af et i forvejen eksisterende ord, fx *retorik*, og 208 ord er nye ord i form og indhold (fx *klangmassage* og *straksdom*).

Ordtrawleren finder en del nye ord som også blev fundet i guldstandarden, men den finder også ord som i en resultatliste ville være støj. Det er bøjningsformer af ord, fx *inuitterne* (*inuit* står i Retskrivningsordbogen), fx fagsprog som *vasopressin* (medicinsk fagsprog), gamle ord som kan stå i Ordbog over det Danske Sprog, men ikke forekommer i andre ordbøger, fx *fiskeplads* og *vurderingspris*. Men den største gruppe af ord som giver støj, er banale sammensætninger som *rejserådgivning*, *rejsevaccination*, eller kometord som *mælkeskandale*. En støjkilde er at maskinen tager fragmenter af flerordsudtryk med som nyordskandidater, fx medtager den *24-års* når det nye ord *24-års-regel* bliver skrevet *24-års regel*, altså uden bindestreg. Institutionsnavne som FEMA, ABX, BRC er også kilde til støj, langt de fleste skal ikke

med, men det er vanskeligt at opstille maskinregler for medtagelse og udelukkelse. Alt i alt er Ordtrawlerens største udfordring håndteringen af banale sammensætninger, fagsprog og institutionsnavne.
I de følgende afsnit vil vi med tørre tal evaluere hvor godt/dårligt Ordtrawleren excerperer i forhold til den menneskelige guldstandard. I eksperimentet har Ordtrawleren anvendt tre forskellige teknikker, og disse teknikker evalueres hver for sig.

Første fremgangsmåde: primitiv filtrering
Som tabel 3 nedenfor viser, så er én tilgang til maskinel excerpering at lade Ordtrawleren fjerne et større antal allerede kendte ordformer fra analysekorpusset. De kendte ordformer kan stamme fra relevante ordbøger og korpusser, og i alt kan ca. 1,2 millioner ordformer elimineres på denne facon. For Den Danske Ordbog og Ordsamlingen var det imidlertid ikke muligt at generere samtlige bøjningsformer automatisk, og derfor det kun lemmaformerne fra disse kilder som tæller med i filter nr. 2 og 3. Det reelle antal ordformer for alle fem filtre er dermed væsentligt højere end angivet i tabellen.

TABEL 3. MASKINEL FILTRERING AF ALLEREDE KENDTE ORDFORMER

Nr.	*Filter*	*Antal lemmaer*	*Antal ordformer*
1	Retskrivningsordbogen 2001	64.038	399.062
2	I Den Danske Ordbog, men ikke i 1	34.960	? (lemmaer anvendes)
3	I Ordsamlingen (sep. 2008), men ikke i 1-2	221.679	? (lemmaer anvendes)
4	I Korpus 90, men ikke i 1-3	?	124.585
5	I Korpus 2000, men ikke i 1-4	?	436.004
I alt		?	**1.216.290**

TABEL 4. MASKINEL KONTRA MANUEL EXCERPERING

	Antal nyordskandidater	*succesrate*	*genkaldelsesrate*	*træfrate*
Menneske	208	1	100 %	100 %
Maskine (inklusive proprier)	1061	0,22	69 %	13 %
Maskine (eksklusive proprier)	589	**0,31**	60 %	**21 %**
Maskine (inklusive proprier og uden Korpus 2000 som filter)	1498	0,20	**84 %**	12 %
Maskine (eksklusive proprier og uden Korpus 2000 som filter)	878	0,28	73 %	17 %

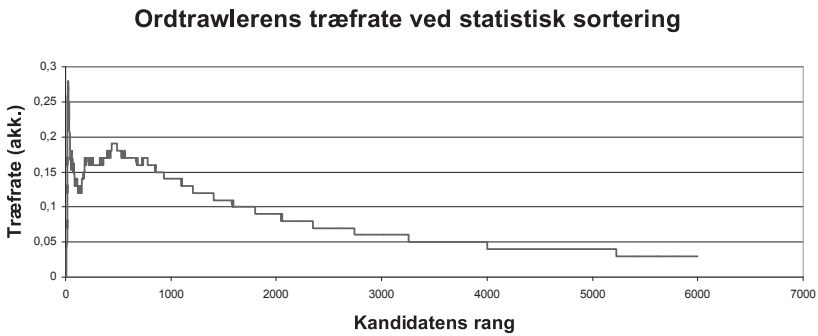
Tabel 4 viser hvordan maskinen klarer sig i forhold til den menneskelige excerpist (som vi antager er 100 % perfekt). De tre mål, *precision*, *recall* og *F-score,* stammer som sagt fra forskningsfeltet *Information Retrieval* og kan på dansk benævnes som henholdsvis træfrate, genkaldelsesrate og succesrate, hvilket vi vil gøre i resten af denne artikel.

Den bedste balance mellem træfrate og genkaldelsesrate opnås ved at anvende samtlige filtre og samtidig udelukke alle proprier i materialet (589 kandidater hvoraf 124, eller 21 %, er korrekte). Det fremgår også at filtrering med alle ordformer i Korpus 2000 eliminerer en del gode kandidater (dvs. reducerer genkaldelsesraten), men samtidig har en vis støjreducerende effekt. Selvom Korpus 2000 repræsenterer tekster fra 1998-2002, kan det altså sagtens indeholde ord som stadig kan betragtes som relativt nye i dag.

Anden fremgangsmåde: statistisk sortering

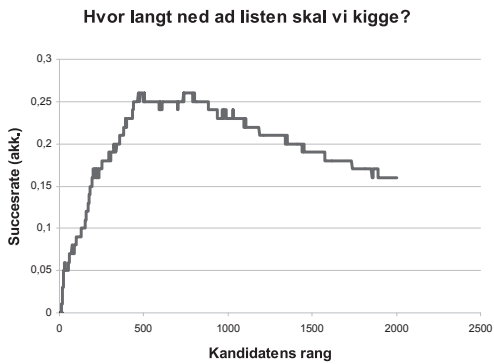
En lidt mere sofistikeret tilgang er at beholde alle ordformer i analysekorpusset, men sortere dem statistisk ved at sammenligne hver ordforms hyppighed i analysekorpusset med dens hyppighed i et stort referencekorpus af ældre dato. Det resulterer i et statistisk mål[24] for hvor bemærkelsesværdigt over- eller underrepræsenteret hver ordform i analysekorpusset er i forhold til referencekorpusset. Hypotesen er at ordformer som er bemærkelsesværdigt overrepræsenterede i analysekorpusset (i dette eksperiment: artiklerne fra Jyllandsposten) i forhold til referencekorpusset (Korpus 2000) er mulige nydannelser.

FIGUR 3. STATISTISK SORTERING AF NYORDSKANDIDATER: TRÆFRATEN.



Figur 3 ovenfor afbilder den akkumulerede træfrate[25] som en funktion af nyordskandidatens rang (dvs. position på resultatlisten). Det fremgår at træfraten ved statistisk sortering på intet tidspunkt kommer over 30 % (0,3).

FIGUR 4. STATISTISK SORTERING AF NYORDSKANDIDATER: SUCCESRATEN.



Én ting er hvor meget "støj" der er på den liste nyordskandidater som Ordtrawleren frembringer, en anden ting er hvor stor en del af samtlige gode kandidater i materialet systemet kan finde. Balancen mellem de to succeskriterier er som sagt succesraten, og figur 4 viser at den bedste balance opnås ved at tage de øverste godt 500 nyordskandidater i betragtning. Dette tal passer meget godt med det antal kandidater der er tilbage efter primitiv filtrering hvor proprier udelukkes (se tabel 4).

Statistisk sortering har altså, i dette eksperiment, næsten samme effekt som primitiv filtrering, men samtidig den fordel at nye betydninger af eksisterende ordformer ikke udelukkes (og ved større tekstmængder vil statistisk sortering være den mest attraktive løsning).

Tredje fremgangsmåde: Nyhedsmarkeringer i konteksten

En åbenlys svaghed ved de to ovenstående fremgangsmåder er at de ser helt isoleret på de enkelte ordformer i teksten og ignorerer konteksten, selvom denne ofte kan indeholde vigtige tegn på at der er en sproglig nydannelse i farvandet. En meget simpel teknik er således at lede efter et antal konkrete nyhedsmarkeringer og excerpere de ord som optræder ved siden af disse signaler.

Vi har forsøgsvis anvendt signalerne "såkaldt", "såkaldte" og citationstegn og bedt maskinen excerpere ordet umiddelbart til højre for de to førstnævnte signaler samt ord der er forsynet med citationstegn. Det resulterede i en fangst på 15 kandidater hvoraf de 6 var korrekte nydannelser ifølge guldstandarden. Eksperimentet viser at nyhedsmarkeringer medfører en høj grad af træfsikkerhed, men en lav genkaldelsesrate (rigtigt mange udmærkede kandidater overses fordi nyhedsmarkeringer er relativt sjældne). Nyhedsmarkeringer kan altså ikke stå alene og kræver desuden store mængder tekst, men informationssamfundet er jo netop karakteriseret ved en eksponentielt stigende tekstproduktion.

Konklusion

Efter vanlig målestok i *Information Retrieval* er vores træfrate (og succesrate) skuffende lav, og den viser at opgaven (excerpering af nye ord til en nyordsordbog) er overordentlig svær for en maskine. Med ca. én fuldtræffer for hvert femte ord vil toppen af Ordtrawlerens kandidatliste dog trods alt være anvendelig og udgøre en tidsbesparelse for den menneskelige excerpist. Samtidig er filtrering via nyhedsmarkeringer en meget lovende teknik når der er tale om store mængder tekst, for så er det mindre afgørende at optimere genkaldelsesraten. Derfor slippes Ordtrawleren i det følgende eksperiment løs på millioner af løbende ord og anvender nyhedsmarkeringer som excerperingsteknik.

2. EKSPERIMENT: EVALUERING AF TRÆFRATE PÅ STORT KORPUS

Til forskel fra det første eksperiment evalueres her alene de af systemet fremfundne nyordskandidater. Der er altså ingen evaluering af hvad analysekorpusset ellers måtte indeholde af genuine nydannelser og dermed af hvor meget systemet "overser" (dvs. *silence*).

Empirien er 96,7 mio. løbende ord fra kortere nyhedsartikler i 55 forskellige danske dagblade i perioden 9. oktober 2007 til 11. oktober 2008. Baseret på resultaterne af evalueringen af artiklens 1. eksperiment valgte vi at lade Ordtrawleren anvende en meget restriktiv excerperingsteknik som kombinerede den primitive filtrering med nyhedsmarkeringer i konteksten. Med andre ord skal alle nyordskandidater være hidtil usete ordformer som kollokerer med mindst én nyhedsmarkering og forekommer mindst to gange i analysekorpusset.

Denne teknik resulterede i 1784 nyordskandidater hvorfra der blev udtrukket de 200 mest frekvente og de 200 mindst frekvente kandidater. Den implicitte hypotese var at høj frekvens ville være en indikation på høj nyordsværdi, og med denne fremgangsmåde var det muligt at undersøge om frekvens kunne fungere som relevansparameter.

Træfrate

Artiklens to forfattere fungerede som evaluatorer i dette eksperiment, og ud af de 400 nyordskandidater blev 152 (af begge evaluatorer) evalueret som relevante nok til at indgå i Ordsamlingen og dermed på længere sigt sandsynligvis også i en nyordsordbog. En træfrate på knap 40% er væsentligt bedre end hvad Ordtrawleren kunne formå med primitiv filtrering og statistisk sortering i artiklens 1. eksperiment, og svarer til resultatet af et lille piloteksperiment hvor kollokerende nyhedsmarkeringer identificerede 6 korrekte nydannelser ud af 15 kandidater i dette eksperiments analysekorpus (artiklerne fra Jyllandsposten). Piloteksperimentet indikerer imidlertid at denne høje træfrate naturligvis opnås på bekostning af en kraftigt reduceret (men ikke praktisk målbar) genkaldelsesrate. Som tidligere nævnt bør nyhedsmarkeringer derfor ikke stå alene, og man skal også være klar over at visse nyhedsmarkeringer (fx "såkaldt(e)") kun kan fremfinde visse konstruktioner (fx NP'er).

Intersubjektiv analyse

De to evaluatorer vurderede at henholdsvis 173 og 157 nyordskandidater var korrekte. Foreningsmængden udgjorde 180 kandidater og fællesmængden 152. Der var med andre ord enighed i 152 ud af 180 tilfælde, hvilket giver en enighedsgrad på 84,4 %.

Evaluering af støj

Tabel 5 viser de forskellige typer af støj der blev observeret i systemoutput samt fordelingen på de forskellige støjtyper. Den "støj" der behandles, er de 400 nyordskandidater fra 2. eksperiment fratrukket de 152 kandidater som blev evalueret som relevante. Hver kandidat kan sagtens repræsentere mere end én støjkategori. Fx kan der være tale om en fagsproglig term som samtidig er stavet forkert.

TABEL 5. ORDTRAWLERENS STØJ.

Støjtype	***Eksempel***	***Antal***	***Andel***
Bøjningsform	*Enhedslønomkostninger, undersøgelseskommissioner*	83	28,8 %
Banale sammensætninger og lejlighedsdannelser	*forskningskvalitet, forårsprognose, fodboldekspert, pizzabande*, *havnepulje*	81	28,1 %
Fagsprog	*FISH-metode, kapillærvirkning*	44	15,3 %
Stavefejl	*billediagnostiske, denial-of-service-angreb*	25	8,7 %
Filterfejl	*nummerportering, artmoney*	14	4,9 %
Kodeskift	*Surge, caucus, caviats*, *stâuerna*	13	4,5 %
Proprium	*JPMorgan, SEA-Games, TMM*	12	4,2 %
Gammel	*epidemihus, dyppelys*	11	3,8 %
NP-fragment	*(§) 20-spørgsmål, hospitality (manager), parkér (og rejs)*	5	1,7 %
I alt		288	

Bøjningsformen som støjkategori fremkommer ved at Ordtrawleren registrerer som nyt ord en bøjningsform som i grundformen allerede er i Ordsamlingen. På baggrund af tallene i tabel 5 er det tydeligt at bøjede former af allerede kendte ord er den støjtype som volder Ordtrawleren de største problemer. Årsagen er naturligvis den manglende lemmatisering af ukendte ordformer (altså ordformer som ikke kan lemmatiseres automatisk ved hjælp af Retskrivningsordbogen 2001).

De næstmest problematiske støjtyper er banale sammensætninger og lejlighedsdannelser. Disse to typer har det vist sig at være vanskelige at adskille, og de er også nogle gange vanskelige at skelne fra de nye, blivende ord. Klare lejlighedsdannelser er *kommunedans*, *pizzabande*, *sponsorsag*, *2015-mål*, *2015-plan*, mens flere af de ord som vi selv i evalueringen bedømte forskelligt, er vanskelige at afgøre endeligt. Det er ord som *biomassebekendtgørelse*, *burmarubin*, *designmaleri*, *enkeltmandskontor*, *højhastighedsskinne*. I en manuel excerpering ville man undersøge disse ord nærmere for at afgøre om de skal excerperes og senere med i en nyordsbog. Det vil først og fremmest være at undersøge ordets udbredelse i tid og genre, at undersøge deres betydning og leksikaliseringsgrad, og om de findes i de ordbøger som ikke indgår i Ordtrawlerens filtre, i andre tekstkorpusser og i leksikoner og specialordbøger. Disse ordtyper vil desværre være ganske vanskelige at reducere maskinelt. Selv med en automatisk opløsningsalgoritme for sammensatte konstruktioner, så er semantikken mellem de enkelte led ikke triviel at analysere. Her er det eneste håb sandsynligvis en diakron analyse som beskrevet i afsnittet om reduktion af støj nedenfor.

At skelne fagsprog fra almensprog er i sig selv en vanskelig opgave (også for et menneske), men opgaven bliver særlig svær for en maskine, når der er tale om at skelne ekstremt sjældne almensproglige ord fra fagsproglige udtryk (som optræder sjældent i almensprog). Igen kan en diakron analyse måske være en hjælp her.

Stavefejl registreres som en nyhed, men er det naturligvis ikke. De fleste, ja hele 65 % af fejlene er fejl i brugen af bindestreg, fx *exit-poll*, *denial-of-service-angreb*, *nul-tolerence-politik*, *nul-tolerancepolitik*.

Filterfejl er fejl i Ordtrawleren som burde have filtreret ordet fra fordi det findes fx i Ordsamlingen eller i et af de andre filtre, men det

er altså ikke sket. Det skal undersøges nærmere hvorfor, og selvfølgelig rettes i næste version af Ordtrawleren.

Ordtrawleren finder sommetider ord som er gamle. Det kan være ord opført i ældre referenceværker som ikke indgår i systemets filtre. Ordet *dyppelys* står således i Ordbog over det Danske Sprog, og godt nok om lys, men her er det karakteriseret som forældet eller dialekt. Ordet er således et eksempel på at Ordtrawleren finder et gammelt ord i ny brug. Ordet *hjælpearbejde* er med i Ordbog over det Dansk Sprogs Supplement, bind 4.

Ordtrawleren har i nogle tilfælde kun genkendt en del af en længere frase, fx *20-spørgsmål* svarende til *paragraf 20-spørgsmål*. Det skyldes at retskrivningsnormen foreskriver at denne slags gruppesammensætninger skal skrives med kun én bindestreg, og at der mangler syntaktisk analyse i systemet.

PINGVINORD

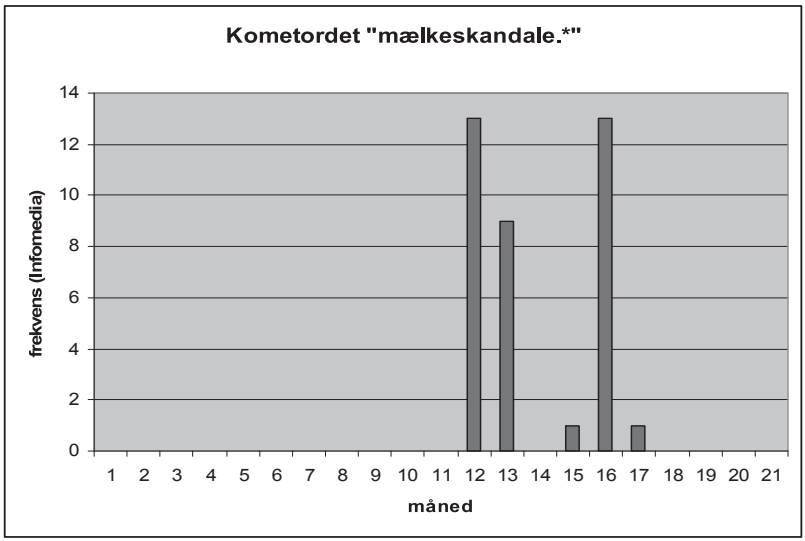
Under redaktionen af Dansk Sprognævns Retskrivningsordbog, 1986, fremlagde en af redaktørerne sin prøveredaktion af bogstavet *P* – og ordet *pingvin* manglede. Siden blev pingvinord betegnelsen for helt almindelige ord som af uransagelige årsager overses af mennesket – af excerpisten, redaktøren osv. Pingvinordene er ikke helt ualmindelige, og det kan hænge sammen med at man fokuserer på ét ord og overser det der står lige ved siden af. *Persillesovs*, som næsten må siges at høre til dansk kulturarv, er først opdaget manglende i Retskrivningsordbogen i forbindelse med arbejdet med den kommende udgave[26].

Ordet *næsehjul* 'støttehjul til campingvogn, fly o.l.' findes af Ordtrawleren, og her har vi at gøre med et pingvinord; *næsehjul* er ikke med i Ordbog over det Danske Sprog, Den Danske Ordbog, Retskrivningsordbogen eller Nudansk Ordbog. Og det er heller ikke nyt. I John Foltmann: Flyveordbogen, 1945 er ordet *næsehjul* med i denne betydning: 'forreste understelshjul'. Med hensyn til pingvinord giver Ordtrawleren overraskende og meget nyttig information. Som human excerpist kan man undre sig over at disse ord er overset: *billigcigaret*, *F1-lån*, *forsvarsforbehold*, *klubånd*, *talentkonkurrence*.

Reduktion af støj

Støjanalysen viste at der med fordel kan sættes ind over for problemerne med henholdsvis bøjningsformer og lejlighedsdannelser, herunder kometord. I dette afsnit vil vi især fokusere på sidstnævnte problem og beskrive de støjreducerende muligheder i diakrone frekvensprofiler, dvs. en afbildning af den hyppighed hvormed en given nyordskandidat forekommer i et givet korpus som repræsenterer et antal sekventielle tidsperioder. I artiklens 1. eksperiment har vi kun empiri for en enkelt dag (den 20. september 2008), men er det muligt at identificere kometord ved at analysere deres frekvensprofil i Sprognævnets diakrone korpus fra Infomedia[27]? Figur 5 illustrerer hvordan en enkelt nyordskandidat fra eksperimentet, nemlig den trunkerede streng mælkeskandale*, afslører sig selv som et kometord via en sådan frekvensprofil.

FIGUR 5. EN DIAKRON FREKVENSPROFIL.



Mælkeskandale observeres første gang i måned nummer 12 (dvs. september 2008) og sidste gang i måned nummer 17 (dvs. februar 2009). Vi planlægger nærmere studier af disse diakrone frekvensprofiler for at se om de kan anvendes som støjfilter.

Et andet oplagt støjreducerende tiltag ville være at reducere stavefejl ved at sammenholde tekststrenge med bindestreg og uden bindestreg. Ordtrawlerens filterfejl skal som nævnt undersøges nærmere, og selvfølgelig rettes i næste udgave. Der er formodning om at de skyldes typer af citationstegn som Ordtrawleren betragter som bogstavtegn. Endvidere er det en oplagt forbedringsmulighed at udruste Ordtrawleren med en lemmatiseringsalgoritme som kan gætte på opslagsformen af ukendte ordformer, også selvom intet programmel vil være ufejlbarligt til denne opgave. Endelig har Ordtrawleren problemer med gruppesammensætninger. Dem er der dog ingen grund til at gøre så meget ved eftersom retskrivningsnormen antages at ændre sig på dette punkt.

Højfrekvente kontra lavfrekvente kandidater

De 400 nyordskandidater i artiklens andet eksperiment repræsenterer som sagt de 200 mest frekvente og 200 mindst frekvente kandidater i det samlede materiale (dvs. de 1784 kandidater som kollokerer med mindst én nyhedsmarkering). Den implicitte hypotese var at høj frekvens ville være en indikation på høj nyordsværdi, men en optælling viser mærkværdigvis ingen nævneværdig forskel på nyhedsværdien af højfrekvente kontra lavfrekvente kandidater (henholdsvis 78 kontra 74 træffere).

At der er mange neologismer blandt lavfrekvente ordformer i et vilkårligt korpus kan virke overraskende, men fænomenet er også blevet påvist i APRIL-projektet (Renouf 2002) hvor der advares mod at ignorere singletoner (dvs. ordformer som kun forekommer én gang i korpus), da disse ikke behøver at være tastefejl eller kometord. Rå frekvens kombineret med nyhedsmarkeringer er dermed en parameter vi vil undersøge nærmere i fremtidige studier.

SAMLET KONKLUSION OG PERSPEKTIVER

Artiklen illustrerer at menneskers og maskiners tilgang til excerpering af sproglige nydannelser fra løbende tekst er vidt forskellig. Den menneskelige excerpist har en række fordele frem for maskinen, især sin fænomenale modersmålskompetence, sin omfattende omverdensviden

og intuition. På alle disse punkter er maskinen underlegen og dens natursprogsbehandling er mangelfuld og fejlbehæftet. Maskinens styrke er til gengæld at den excerperer 100 % objektivt, ikke har fokusproblemer (som fx kan få den til at overse ovennævnte pingvinord) og har en langt hurtigere processeringstid.

Konklusionen er altså at man må fortsætte med at excerpere manuelt, idet en række sproglige nydannelser ikke kan identificeres af Ordtrawleren på nuværende tidspunkt (fx flerordsforbindelser, nye valensmønstre, ny brug af gamle ord, billedlig brug), men at den automatiske excerpering, med den rette filtrering/sortering, giver et væsentligt tilskud til mængden af nyordskandidater. Desuden vil Ordtrawlerens træfrate sandsynligvis øges efterhånden som korpusset spænder over længere tid, og lejlighedsdannelser og kometord bedre kan identificeres.

Jakob Halskov
Dansk Sprognævn
jhalskov@dsn.dk

Pia Jarvad
Dansk Sprognævn
jarvad@dsn.dk

NOTER

1 http://www.nyeordidansk.dk.

2 Lov om Dansk Sprognævn, Lov nr. 320 af 14. maj 1997.

3 Dette år er forbilledligt i fordeling af kilder. Det afspejler dels at arbejdet med Nye ord i dansk 1955-75 (Jarvad 1984) viste at excerperingen var skæv, dels at Arne Hamburger (dagbladet Information) og Jørgen Eriksen (dagbladet Politiken) gennem mange år excerperede de to dagblade så tæt at andre kilder slet ikke kunne slå igennem, og selve det fysiske arbejde med at overføre excerptet fra streg til samling gjorde at mindre hyppigt excerperede kilder var meget lang tid om at komme i samlingen og dermed være søgbare. Derfor blev excerperingen lagt i skema, med begrænsning af antallet af excerpter fra Information og Politiken og fokusering på andre kilder. Men det holdt dog ikke i længden, lystexcerpister er ikke nemme at styre.

4 I 2007 blev en ideel plan for excerpering opstillet, hvor de forskellige genrer, tekster mv. blev vægtet efter bl.a. viden om tekstartens udbredelse, og om hvor man finder nye ord.

5 Tallet dækker over både provins- og hovedstadsaviser.

6 Provinsaviser erkendes som objekt for excerpering.

7 Distriktsblade er husstandsomdelte blade, og de dækker både provins og hovedstad.

8 Der er i disse to år excerperet distriktsblade, men blev i opgørelsen henregnet til ugeblade.

9 Kortgenre er excerpter fra skilte, reklamer, hvor kontekst ofte er lille eller helt mangler. Ofte dokumenteret med fotos, emballage mv.

10 Den nye generation af ansattes ønsker om mindre finkulturel excerpering slår nu igennem. Når så denne kilde stiger senere, skyldes det at også tegneserier og ungdomsblade nu anses for at være givtige for excerpering.

11 Ikke decideret faglitteratur, men derimod mere alment faglige som fx Helse, Råd & Resultater, brugsanvisninger til boremaskiner og foodprocessorer.

12 Officielt sprog er en vigtig kilde både for Sprognævnets forpligtelse over for varetagelse af juridisk sprog som et klart og gennemskueligt sprog og for vurderingen af et nyt ords status.

13 Litteraturen var for Ordbog over det Danske Sprog den vigtigste kilde; moderne litteratur er ikke mere det vigtige sted for varige nydannelser, jf. SiN 1976. Bøger står her for skønlitteratur.

14 Fremmedsprog excerperes for at dokumentere etymologi og de europæiske sprogs fælles udvikling. Her er i særlig grad tale om The Guardian, Le Monde, Neue Züricher Zeitung.

15 http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=GR/L08243/01.

16 http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

17 http://www.natcorp.ox.ac.uk/.

18 http://korpus.dsl.dk/e-resurser/k90_info.php?lang=dk.

19 En delsekvens af en længere sekvens af enheder, i dette tilfælde ord.

20 http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13.

21 Delsekvenser på én enhed, fx "manden".

22 Delsekvenser på to enheder, fx "manden som".

23 Tak til Anne Sofie Jakobsen.

24 Vi anvender målet *log-odds* (jf. Evert 2004) som i særlig grad prioriterer sjældne begivenheder (fx ordformer som aldrig er set før).

25 Dvs. den samlede træfrate for alle nyordskandidater frem til og med den givne rang, fx de øverste 10 eller 100 kandidater.

26 Tak til Jørgen Schack for eksemplet.

27 Dette korpus indeholder pt. ca. 21 måneders nyhedstekster fra og med oktober 2007. I alt ca. 150 mio. løbende ord.

LITTERATUR

Atkins, B. T. Sue & Michael Rundell (2008) *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Biber, Douglas & Susan Conrad & Randi Reppen (1998) *Corpus Linguistics*. Cambridge: Cambridge University Press.

Evert, Stefan (2004) *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.d.-afhandling, Stuttgart Universitet.

Halskov, Jakob & Pia Jarvad (2009) "Om menneskers og maskiners tilgang til excerpering af sproglige nydannelser – en diskussion og en systemevaluering". *Nyt fra Sprognævnet* 2009/4. København: Dansk Sprognævn.

Halskov, Jakob & Pia Jarvad (2010) "Human versus automated extraction of neologisms for lexicography - a discussion and a system evaluation". *Cahiers du Cental*, vol. 6, Louvain-La-Neuve: Presses universitaires de Louvain.

Hansen, Dorte Haltrup (2000) *Træning og brug af Brill-taggeren på danske tekster*. Teknisk rapport fra Center for Sprogteknologi (CST). http://cst.dk/online/pos_tagger/Brill_tagger.pdf.

Jarvad, Pia (1995) *Nye ord – hvorfor og hvordan?* København: Gyldendal.

Keson, Britt (1998) *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus*. Teknisk rapport fra Det Danske Sprog- og Litteraturselskab (DSL). http://korpus.dsl.dk/paroledoc_dk.pdf.

Kilgarriff, A. & G. Grefenstette (2003) "Introduction to the special issue on the web as corpus". *Computational Linguistics*, vol. 29, nr. 3. Boston: MIT Press. 333-347.

Kilgarriff, Adam & Pavel Rychly & Pavel Smrz & David Tugwell (2004) "The Sketch Engine". Geoffrey Williams & Sandra Vessier (red.) *Proceedings of the Eleventh EURALEX International Conference*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

McEnery, Tony & Andrew Wilson (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Renouf, A. (2002) "The Time Dimension in Modern Corpus Linguistics". Bernhard Kettemann & Georg Marko (red.) *Teaching and Learning by Doing Corpus Analysis. Papers from the 4th International Conference on Teaching and Learning Corpora*. Amsterdam/Atlanta: Rodopi. 27-41.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at work*. Amsterdam: John Benjamins.

Zipf, George K. (1935) *The Psychobiology of Language*. Boston: Houghton-Mifflin.