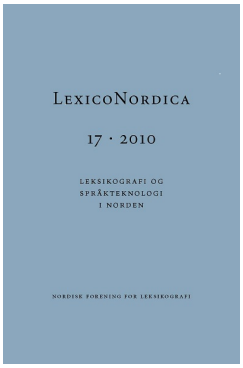
LexicoNordica

Titel: Halvautomatisk udvælgelse af lemmakandidater til en nyordsordbog

Forfatter: Jakob Halskov

Kilde: LexicoNordica 17, 2010, s.73-97

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



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

Halvautomatisk udvælgelse af lemmakandidater til en nyordsordbog

Jakob Halskov

One of the key tasks of the Danish Language Council (Dansk Sprognævn) is to monitor, record and document linguistic changes in the Danish language. To assist in this task, a prototype neologism detector called the *Ordtrawler* (Word trawler) is being developed. The system processes large amounts of text and extracts candidate neologisms using a combination of simple filters, collocational statistics and so-called neology markers. A thorough evaluation of the *Ordtrawler* indicates that neology markers are good for optimizing system precision, but at the expense of recall. Certain types of noise such as technical terms and semantically transparent compounds remain to be tackled, but diachronic frequency profiling may help.

1. Indledning

Denne artikel beskriver arbejdet med Dansk Sprognævns "Ordtrawler", en softwareprototype som har til formål automatisk at finde nyordskandidater i de meget store danske tekstkorpusser som i dag er tilgængelige på nettet. Der er foretaget en formel evaluering af dette automatiske excerperingssystem (se Halskov og Jarvad 2010), men i denne artikel er der fokus på de mere kvalitative aspekter af programmellet. Det beskrives således hvilke typer nydannelser systemet i øjeblikket kan identificere (afsnit 2), hvordan de identificeres (afsnit 3), og hvilke typer nydannelser det er mere vanskeligt at håndtere maskinelt. Endelig skitseres det hvordan systemet kan videreudvikles til at håndtere mere "avancerede" typer nydannelser ved eksempelvis at trække på nye sprogteknologiske resurser som DanNet-ontologien (afsnit 5).

1.1. Hvad er et nyt ord?

Sproglige nydannelser kan forekomme på alle lingvistiske niveauer, lige fra det semantiske niveau over de fraseologiske og syntaktiske niveauer til det ortografiske niveau. I denne sammenhæng er det primært det (leksikalsk) semantiske niveau som interesserer os, og på dette niveau kan sproglige nydannelser inddeles i de følgende tre kategorier:

1. Nye ord som refererer til nyt indhold (fx *app* om software til moderne mobiltelefoner)
2. Eksisterende ord som får nyt indhold (fx *skyde* i betydningen '(op)tage' om film)
3. Eksisterende ord som erstatter eksisterende indhold (fx *omkring* for *om*)

Kategori 1 og 2 kaldes også henholdsvis neoformativer og neosemantismer (Fjeld og Nygaard 2010). Kategori 2 og 3 kan samordnes under betegnelsen "ny brug", og denne brede betegnelse kan også omfatte ny valens eller ændrede selektionsrestriktioner og dermed involvere det syntaktiske niveau. Kategori 1 kan naturligvis også omfatte nye flerordsudtryk eller neofrasemer (Fjeld og Nygaard 2010).

Nye ord og ny brug af gamle ord er nyheder i forhold til det inventar af ord og betydninger som findes i forvejen. Mængden af opslagsord og betydninger i gængse ordbøger[1] udgør naturligvis kun en delmængde af det samlede danske almensproglige ordforråd. "Ny" bruges derfor i denne kontekst i betydningen ny for leksikografen og det leksikografiske miljø.

1 Fx Den Danske Ordbog, Retskrivningsordbogen, Nudansk Ordbog, og ikke mindst Nye ord i dansk på nettet fra 1955 til i dag (www.nyeordidansk.dk) og Dansk Sprognævns Samling.

1.2. Et spørgsmål om prægnans

Nydannelser som ikke antages at ville blive etablerede i ordforrådet, har ingen varig effekt på sproget og er dermed ikke interessante som lemmaemner i forhold til en nyordsordbog. Selvom det er vanskeligt at spå om nydannelsers fremtid, så er der kategorier af nydannelser som det, for den menneskelige excerpist, er relativt let at afvise som lemmakandidater, nemlig banale sammensætninger, lejlighedsdannelser og, i et vist omfang, "kometord" (Jarvad 1995:173).

Banale sammensætninger er fx *klimakonference*, *værdipapirsammensætning* og *risikoappetit*. Orddannelsen følger de grammatiske regler, og resultatet er semantisk transparente sammensatte ord som er helt uproblematiske at forstå hvis man kender førsteled og sidsteled. Banale sammensætninger kan ofte være relevante at indlemme i "almindelige" mono- og bilingvale ordbøger, men da de ikke har nogen nyhedsværdi, bør de ikke indlemmes i en nyordsordbog. Indlemmede man dem, ville nyordsordbogens lemmainventar hurtigt antage astronomiske dimensioner (jf. figur 1).

Lejlighedsdannelser er ofte knap så gennemskuelige, idet deres semantiske indhold er meget afhængigt af den kontekst hvori de er ytret. Til gengæld er de, for redaktøren af en nyordsordbog, lige så irrelevante som de banale sammensætninger på grund af deres forbigående karakter. Jarvad (1995) har følgende betragtninger om lejlighedsdannelser:

> Vi kan lave ord som har et mere tilfældigt præg, fx *tefest* (jf. *vinfest*), *tesøster* (jf. *kaffesøster*), *tetår* (jf. *kaffetår*), og *tetelt* (jf. *øltelt*). [...] Disse ord kaldes øjebliksdannelser eller lejlighedsdannelser, og nogle kalder dem individualdannelser. Om disse ords forståelighed er der det at bemærke at de måske nok forstås, men det bagvedliggende ord som fx *kaffesøster*, *øltelt* med disse ords bibetydninger gør forståeligheden større. (Jarvad 1995:173)

En helt særlig udfordring udgøres af ord som man umiddelbart ikke ville spå nogen særlig levetid, men som pludselig går hen og bliver meget udbredte i sprogsamfundet i en kortere periode hvorpå de så forsvinder igen. De ord som forsvinder igen, kan retrospektivt kaldes "kometord". Nogle nyere eksempler på kategorien er *burkaudvalg, klimakaravane* og *mælkeskandale*. Kometord får per definition aldrig en central position i ordforrådet, for hvis de gjorde, ville deres frekvensprofil ikke være kometagtig, altså pludseligt og kraftigt stigende fra nul og derpå hurtigt aftagende til nul igen. Alligevel kan nyordsleksikografen godt komme til at indlemme kometord i sin nyordsordbog eller udelade prægnante lemmakandidater grundet en fejlagtig antagelse om lemmaets kometstatus. Det skyldes naturligvis at det at verificere en kometagtig frekvensprofil kan fordre en tidshorisont som skal måles i år snarere end måneder.

Når lemmakandidater skal selekteres til en nyordsordbog, er den afgørende kvalitetsparameter deres anslåede prægnans, altså med hvilken sandsynlighed de kan blive varige tilskud. Som sagt, og som understreget i Fjeld og Nygaard (2010), så kræver det derfor en vis udbredelse over en vis tid at skelne prægnante nydannelser fra kometord.

> Alle nyord kan oppfattes som potensielle okkasjonalismer. [...] Bare om de bliver brukt av mange og over tid, regnes de som nyord, som seinere igjen kan bli allmennord. (Fjeld og Nygaard 2010)

Da nyordsordbøger i dag også udgives online, så er det imidlertid muligt løbende at revidere ordbogens lemmainventar og retrospektivt markere eventuelle kometord som sådanne eller helt fjerne dem. Ordtrawleren fokuserer således i første omgang på at eliminere andre former for støj, herunder de banale sammensætninger, og lader håndteringen af kometordene indgå i det fremtidige udviklingsarbejde (se afsnit 5).

2. Hvordan excerperer en maskine?

De tekstkilder hvorfra der ved Dansk Sprognævn manuelt er blevet excerperet nye ord siden 1955, omfatter aviser, ugeblade, tidsskrifter, bøger og mange andre typer tekster som repræsenterer forskellige genrer. Den halvautomatiske excerpering, som foretages af Ordtrawleren, har indtil videre taget udgangspunkt i et større antal dagblade der er tilgængelige i elektronisk form via Danmarks største artikeldatabase, InfoMedia. InfoMedia indeholder p.t. ca. 20,6 millioner avisartikler. Der er dog planer om at udvide den halvautomatiske excerpering til også at omfatte en række nyere elektroniske medier, såsom chat, blogs, Facebook og lignende.

Mens den menneskelige excerpist kan trække på sin ekspertviden om modersmålets orddannelse og ordforråd og fx anvende et omfattende antal referenceværker til at tilbagedatere nyordskandidater, så er den maskinelle excerpist på én gang mennesket underlegen og overlegen.

Underlegenheden skyldes primært at den automatiske natursprogsbehandling slet ikke kan konkurrere med menneskets. Mennesket lemmatiserer og normaliserer problemfrit enhver nyordskandidat uanset om kandidaten indgår i dets aktive ordforråd eller ej, men maskinen skal have detaljerede instrukser om enhver form for sproglig analyse og er kun i begrænset omfang i stand til at gætte på forhold vedrørende ord som den ikke har i sine ordbøger. Maskinen har stadig svært ved at abstrahere medmindre den er blevet eksplicit programmeret til det. Derfor vil den som udgangspunkt opfatte stavevarianter som vidt forskellige ord og fx foreslå *U.S.A.* som nyordskandidat selvom den har formen *USA* i sin liste over allerede kendte ord. På tilsvarende vis skal en maskine 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 inddra-

ge 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 (mere herom i afsnit 5).

Overlegenheden består derimod primært i at maskinen ikke har problemer med fokusere. Den distraheres ikke af en succesoplevelse, er ikke særligt interesseret i nydannelser inden for bestemte faglige domæner eller på bestemte lingvistiske niveauer, og endelig kan den bearbejde enorme mængder empiri på kort tid og 100 % systematisk.

2.1. Hvilke excerperingssystemer findes der?

Der findes adskillige natursprogsbehandlingssystemer til automatisk detektion af termkandidater (potentielt fagsproglige udtryk) i et tekstkorpus, fx Termostat[2] (Drouin 2003), TermExtractor (Pantel og Lin 2001) og tidlige implementeringer som den beskrevet i Ahmad (1993). Til gengæld er der nærmest ikke publiceret nogen specifikationer for systemer som kan excerpere almensproglige eller fagsproglige nydannelser. Lad det dermed ikke være sagt at sådanne systemer ikke eksisterer, for det engelske APRIL-projekt (A knowledge-rich tool for the analysis and prediction of innovation in the lexicon)[3] er eksempelvis et ret velkendt, omend udokumenteret, et af slagsen. Desuden er der et enkelt helt nyt og nordisk eksempel på et automatisk excerperingssystem som er beskrevet i Fjeld og Nygaard (2010).

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

2 http://olst.ling.umontreal.ca/~drouinp/termostat_web/

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

korpusværktøjer. De anvendte teknikker beskrives i hovedtræk i det følgende afsnit.

2.2. Hvordan excerperer Ordtrawleren?

Som nævnt er Ordtrawlerens primære tekstkilde faste leverancer af korte avisartikler fra InfoMedia. Denne samling af elektroniske avisartikler må underkastes en række automatiserede behandlinger i en bestemt rækkefølge før Ordtrawleren kan excerpere og altså selektere lemmakandidater til en nyordsordbog.

1. **Tokenisering**: Brødteksten deles op i sætninger, og sætningerne hakkes op i en sekvens af ordformer.
2. **Ordklassetagging**: Hver ordform tildeles automatisk en ordklasse (her anvendes de forenklede Parole-tags (Keson 1998)).
3. **Lemmatisering**: Hver ordform tildeles automatisk et lemma.
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*-formatet[4]).
5. **Filtrering/sortering**: Inventaret af samtlige forskellige ordformer filtreres og/eller sorteres ved hjælp af ord- og frekvenslister over allerede kendte ord.

Den automatiske tokenisering, ordklassetagging og lemmatisering er naturligvis ikke fejlfri. Den anvendte tagger 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 taggerens ordbog, volder

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

altså særligt store problemer, og det er i dette tilfælde netop en delmængde af disse ord vi er ude efter. Problemet gælder i særlig grad den lemmatiseringsteknik som i øjeblikket anvendes af Ordtrawleren. Der anvendes nemlig en udfoldet version af Retskrivningsordbogen 2001 hvor samtlige bøjningsformer af hver homograf i alle ordbogsartikler genereres automatisk. Ordformer som ikke kan henføres til en homograf i dette værk, lemmatiseres dermed ikke.

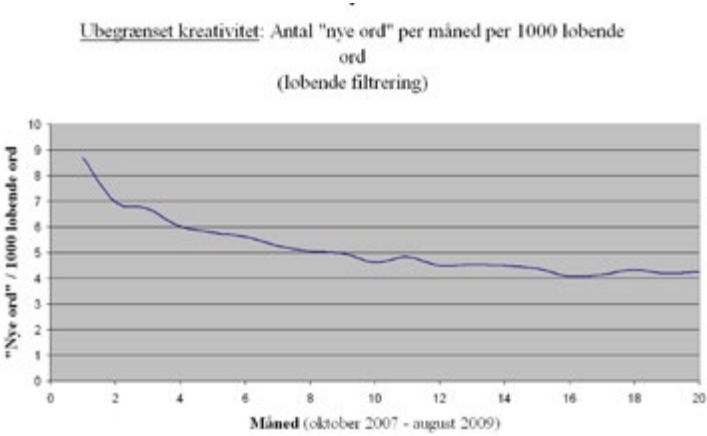
Ordform	*Ordklasse*	*Lemma*
At	UKONJ	at
forbyde	V_INF	forbyde
salg	N	salg
af	PRAEP	af
tobak	N	tobak
er	V_PRES	være
ikke	ADV	ikke
en	PRON_UBST	en
måde	N	måde
at	UKONJ	at
forlænge	V_INF	forlænge
danskernes	N_GEN	dansker
levetid	N	levetid
med	PRAEP	med
.	TEGN	.

Tabel 1: En bearbejdet sætning i det særlige Corpus Workbench-format

Tabel 1 indeholder et eksempel på hvordan en sætning ser ud efter ovenstående automatiserede behandlinger. Når alt tekstmaterialet foreligger i dette format, trækkes hele ordforrådet ud (dvs. alle *forskellige* ordformer, også kaldet "types") og sammenlignes form for form med det ordforråd systemet kender i forvejen. I afsnit 3 beskrives de teknikker systemet anvender til lemmakandidatudvælgelsen, og i afsnit 3.1 beskrives det hvilke eksisterende ordbøger og referenceværker der udgør det kendte ordforråd i denne sammenhæng.

2.3. Ordtrawlerens empiri: Ubegrænset sproglig kreativitet

Dette afsnit illustrerer den grænseløse sproglige kreativitet og produktivitet som Ordtrawleren må forsøge at navigere i. Figur 1 nedenfor viser hvordan Ordtrawleren måned for måned registrerer tusinder af ukendte ordformer i de store mængder nyhedsartikler 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 (fx e-mail-adresser) og proprier og tilmed anvender filtrering af alle hidtil observerede ordformer.



Figur 1: 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 ca. 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 træffe sådanne afgørelser. Tallet kan virke over-

raskende, for der er relativt 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*).

Heldigvis kan man anvende et antal forskellige teknikker til enten at reducere antallet af nyordskandidater betragteligt eller til at sortere i dem på en sådan facon at de bedste kandidater kommer til at rangere højere end de dårligste. Dette er emnet for det følgende afsnit.

3. Ordtrawlerens sprogteknologiske værktøjskasse

I dette afsnit beskrives de tre teknikker som i øjeblikket anvendes af Ordtrawleren til detektion af nyordskandidater fra tekstkorpusser. Det drejer sig om henholdsvis primitiv filtrering, kollokationsstatistik og "nyhedsmarkeringer" (Jarvad 1995:23). I afsnit 4 beskrives det ganske kort hvordan disse tre teknikker med fordel kan kombineres, og hvilke resultater det har givet i en konkret formel evaluering af Ordtrawleren.

3.1. Primitiv filtrering

Selvom det hverken er tilstrækkeligt eller ideelt at fjerne et større antal allerede kendte ordformer fra analysekorpusset, så er det en helt rudimentær teknik som det i det mindste er værd at bruge lidt plads på at diskutere. Fordelen ved primitiv filtrering er at det er en meget simpel teknik som er nem at implementere. Der er dog to store ulemper ved teknikken. Dels frafiltreres der slet ikke nok uprægnante nydannelser og dels elimineres der samtidig en del prægnante nydannelser, eksempelvis ny brug af eksisterende sproglige udtryk.

Tabel 2 viser hvilke primitive filtre der anvendes i den nuvæ-

rende version af Ordtrawleren. For Den Danske Ordbog og Dansk Sprognævns Ordsamling[5] er det kun opslagsordene i deres grundform som har været tilgængelige for Ordtrawleren. Antallet af samtlige bøjningsformer der kan dannes på basis af disse grundformer, er således væsentligt højere og ville bringe den samlede størrelse af de primitive filtre langt over 1,2 mio. ordformer. For Korpus 90 og Korpus 2000 er det omvendte tilfældet. Her er antallet af lemmaer ukendt.

Nr.	Filter	Antal lemmaer	Antal ordformer
1	Retskrivningsordbogen 2001	64.038	399.062
2	I Den Danske Ordbog, men ikke i 1	34.960	34.960
3	I Ordsamlingen (september 2008), men ikke i 1-2	221.679	221.679
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 2: Primitiv filtrering af kendt ordmateriale

3.2. Kollokationsstatistik

En mere sofistikeret tilgang til halvautomatisk nyordsdetektion er at sortere analysekorpussets ordforråd snarere end at filtrere noget fra. Her er kollokationsstatistikken (jf. Dunning 1993) et oplagt valg, idet den med en mindre justering kan måle i hvilken grad en given ordform er særligt over- eller underrepræsenteret i et analysekorpus kontra et referencekorpus og dermed kan siges at være særligt karakteristisk for førstnævnte korpus. Selvom man normalt anvender et såkaldt (firecellet) "contingency table" (Pearson

5 Den elektroniske del af Dansk Sprognævns Ordsamling indeholder p.t. godt 273.000 opslagsord med knap 320.000 sprogbrugseksempler med udførlige kildeangivelser (belæg).

1904) til at sammenholde to ords respektive forekomster med deres samforekomst i ét korpus, kan man anvende samme firecellede tabel til at sammenholde ét bestemt ords respektive forekomst i to forskellige korpusser. De to måder at anvende "contingency tables" på er skitseret i tabel 3.

Baseret på frekvenstallene i de fire celler (kaldet O_{11} til O_{22} i tabel 3) kan man beregne et associationsmål som altså enten udtrykker tiltrækningskraften mellem to ord, fx *rejse* og *penge*, i et givet analysekorpus (jf. fodnote 8) eller mellem ét ord, fx *klimakaravane*, og to forskellige korpusser (jf. fodnote 7 og 8). Eksemplerne i tabel 3 illustrerer at tiltrækningskraften mellem *rejse* og *penge* er stærkere end mellem *rejse* og *over,* og *rejse penge* er dermed en stærkere kollokation end *rejse over*. På samme facon er nyordskandidaten *klimakaravane* mere karakteristisk for analysekorpusset, hvor den forekommer 3 gange, end for referencekorpusset, hvor den har en nulforekomst. Omvendt er *politiker* marginalt mere karakteristisk for referencekorpusset end for analysekorpusset, idet tiltrækningskraften til sidstnævnte er negativ. Tallet ligger imidlertid så tæt på 0 at ordet må siges at være lige karakteristisk for analysekorpusset (2008-2009) som for referencekorpusset (2004-2005).

En række forskellige statistiske mål kan anvendes til at beregne disse associationsstyrker, og forskellene mellem dem beror primært på hvilken vigtighed de tillægger sjældne begivenheder, altså lavfrekvensforekomster. To populære mål er henholdsvis "log-odds ratio"[6] og "log-likelihood ratio". Ordtrawleren anvender i sin nuværende konfiguration primært førstnævnte associationsmål, da dette mål tillægger lavfrekvensforekomster stor vigtighed (jf. Evert 2004). Det engelske APRIL-projekt viste nemlig at der i et vilkårligt korpus typisk vil være mange sproglige nydannelser blandt de mest lavfrekvente ord (Renouf 2002), og dermed passer "log-odds ratio" bedre end "log-likelihood ratio" til formålet.

6 log-odds-ratio = $\log((O_{11}*O_{22})/(O_{12}*O_{21}))$

	ord_2=X	ord_2≠X		ord=X	ord≠X
ord_1=Y	O_{11}	O_{12}	analysekorpus[7]	O_{11}	O_{12}
ord_1≠Y	O_{21}	O_{22}	referencekorpus[8]	O_{21}	O_{22}
Eksempel 1: "rejse penge"			Eksempel 1: "klimakaravane"		
	ord_2= penge	ord_2≠ penge		ord=klima-karavane	ord≠klima-karavane
ord_1=rejse	41	3864-41	analysekorpus	3	68373866-3
ord_1≠rejse	13849-41	39616147-3864-13849	referencekorpus	0	39616147-0
log-odds("rejse penge"): log((41*39598434)/(3823*13808)) **≈ 1,49**			log-odds("klimakaravane"): log((3.5*39616147.5)/(68373866.5*0.5)) **≈ 0.61**		
Eksempel 2: "rejse over"			Eksempel 2: "politiker"		
log-odds("rejse over"): log((13*39531012)/(3851*81249)) **≈ 0.22**			log-odds("politiker"): log(1997.5*39614853.5)/ (68371869.5*1294.5)) **≈ -0.05**		

Tabel 3: Kollokationsstatistik med "contingency tables"

3.3. Potentielle nyhedsmarkeringer

Potentielle nyhedsmarkeringer er eksplicitte sproglige eller typografiske virkemidler som ofte, men ikke altid, ledsager sproglige nydannelser i de første faser af deres liv, altså inden de for alvor måtte blive etablerede i sproget. Det er næppe muligt at give en udtømmende liste over alle potentielle nyhedsmarkeringer, men fænomenet er blandt andet omtalt i *Nye ord – hvorfor og hvordan?* (Jarvad 1995):

> En almindelig udbredelsesmåde for et nyt ord er at det dukker op i en avis, og tingen eller fænomenet bliver omtalt,

7 Dette korpus består af nyhedsartikler fra InfoMedia i perioden oktober 2008 til oktober 2009, i alt ca. 68 mio. løbende ord.

8 En større samling nyhedsartikler fra InfoMedia i perioden december 2004 til oktober 2005. I alt ca. 39,6 mio. løbende ord.

> det bliver forklaret og måske sat i gåseøjne [...] Andre nyhedssignaler er løftede kommaer omkring ordet, ordet kursiveres eller særmærkes grafisk på anden måde [...] (Jarvad 1995:23)

Set fra et sprogteknologisk perspektiv er de tre primære kvalitetskriterier for potentielle nyhedsmarkeringer henholdsvis deres individuelle frekvens og to mål som stammer fra videnskabsgrenen "Information Retrieval", nemlig "precision" (træfrate) og "recall" (genkaldelsesrate). Altså i hvor høj grad identificerer den potentielle nyhedsmarkering sproglige nydannelser og ikke andet, og i hvor høj grad kan markeringen bruges til at finde frem til alle de sproglige nydannelser der måtte være i et givet analysekorpus.

Citationstegn og kursiv er eksempler på potentielle nyhedsmarkeringer med en antaget høj genkaldelsesrate, for de kan nemlig ledsage stort set enhver type sproglige nydannelser (fx "Italien sender nu 'grundløse asylansøgere' tilbage." (Politiken 25.1.2009)). En anden prominent nyhedsmarkering er attributtet "såkaldt(e)" (fx "Meget af denne vold er såkaldt opdragelsesvold." (Politiken 2.11.2009)), men netop denne markering har den ulempe at den ikke kan ledsage VP'er og altså ikke vil kunne anvendes til at identificere nye verber.

Det er vigtigt at understrege at både de typografiske markeringer og de fleste sproglige markeringer, i særdeleshed "såkaldt(e)", naturligvis kan anvendes til at markere en række andre metasproglige forhold i teksten. Eksempelvis en ironisk, politisk eller videnskabssmæssig distance til det skrevne. I disse tilfælde vil markeringen ikke nødvendigvis pege på nydannelser, men ofte på allerede kendt ordmateriale i henholdsvis den almensproglige og den fagsproglige sfære.

Der er en række sproglige markeringer som givetvis vil have en meget høj træfrate (fx "som det hedder på nydansk"), men som grundet lav forekomst i korpus, måske er knap så anvendelige i praksis, jf. tabel 4. Omvendt er der potentielle markeringer der har

en meget høj frekvens, men som ikke i sig selv er tilstrækkeligt præcise, fx bindestreger (se tabel 4). Heidemann (2009) har undersøgt brugen af bindestreg i moderne importord og blandt andet fundet at det primært er svagt leksikaliserede ord og ord der har beholdt deres engelske udtale og skrivemåde, som særskrives. Da brugen af bindestreg er en hybrid mellem den sær- og sammenskrevne form, og da nydannelser som udgangspunkt er svagt leksikaliserede, så giver det god mening forsøgsvis at anvende bindestreg som potentiel nyhedsmarkering. Dette træk kan naturligvis ikke stå alene, men sammen med andre træk kan bindestreger signalere at der er tale om en nydannelse.

Markering	**Frekvens**	**Fremfinder:**	**Antaget effektivitet**[9]
bindestreger[10]	7366 ppm[11]	mest NP'er	Højeste frekvens, moderat genkaldelsesrate, laveste træfrate
citationstegn I (")	832 ppm	alt	Høj frekvens, højeste genkaldelse, lav træfrate
citationstegn II (')	427 ppm	alt	ditto.s.
såkaldt(e)	159 ppm	NP'er	Moderat frekvens, genkaldelse og træfrate
som den|det|de hedder	8 ppm	NP'er, VP'er	Lav frekvens, høj genkaldelse, højere træfrate
som den|det|de kaldes	0,8 ppm	NP'er, VP'er	Meget lav frekvens, høj genkaldelse, høj træfrate

Tabel 4: Potentielle nyhedsmarkeringer og antaget effektivitet

9 Den antagede effektivitet bygger på intuition og erfaring og bør naturligvis undersøges empirisk.

10 Tallet angiver antallet af ord med præcis én bindestreg.

11 Frekvenserne i tabel 4 er baseret på det korpus som er beskrevet i fodnote 6. Ppm står for "parts per million" og angiver i denne kontekst antal forekomster per million løbende ord i korpus.

Som det beskrives i næste afsnit, der omhandler en formel evaluering af Ordtrawlerens evne til at identificere sproglige nydannelser, anvendes der i øjeblikket kun den potentielle nyhedsmarkering "såkaldt(e)", markeret med gråt i tabel 4, da denne antages at repræsentere den bedste balance mellem frekvens, genkaldelsesrate og træfrate. Der er imidlertid planer om at anvende en kombination af potentielle nyhedsmarkeringer, inklusive bindestregerne, og samtidig formelt evaluere hver enkelt markerings træf- og genkaldelsesrate.

4. Formel evaluering af Ordtrawleren

I Halskov og Jarvad (2010) foretages en grundig formel evaluering af Ordtrawlerens evne til at identificere sproglige nydannelser der efter en manuel evaluering, kan optages som lemmaer i en nyordsordbog. Detaljerne i denne evaluering skal ikke gengives her, men resultatet beskrives i dette afsnit i hovedtræk.

Evalueringen omfatter to dele. Først evalueres systemets genkaldelsesrate og træfrate på et mindre analysekorpus på 75.000 løbende ord hvori samtlige sproglige nydannelser var blevet manuelt identificeret af to excerpister, og derpå evalueres træfraten på et stort analysekorpus (96,7 mio. løbende ord).

Den førstnævnte evaluering viste at den primitive filtrering reducerede de ca. 14.000 ordformer i analysekorpusset til 589 nyordskandidater hvoraf 21 % optrådte i guldstandarden, dvs. mængden af nydannelser som de menneskelige excerpister på forhånd havde identificeret i korpusset. Genkaldelsesraten var imidlertid kun 60 % hvilket illustrerer at primitiv filtrering ikke er en optimal teknik, da mange genuine nydannelser elimineres. Evalueringen viste til gengæld at kollokationsstatistikken er en bedre teknik, idet man med denne opnår en tilsvarende træfrate, men uden at filtrere nogen kandidater fra. Endelig blev det påvist at

potentielle nyhedsmarkeringer giver den højeste træfrate (ca. 40 %), men en meget lav genkaldelsesrate.

I evalueringens anden del, hvor fokus var optimering af træfrate uden hensyntagen til genkaldelsesrate, anvendtes en kombination af primitiv filtrering med de potentielle nyhedsmarkeringer. Med denne kombinerede teknik kunne der udtrækkes ca. 1800 nyordskandidater fra det store analysekorpus med en træfrate på knap 40 % og en enighedsgrad mellem de to menneskelige evaluatorer på 84,4 %. Denne træfrate er sammenlignelig med de godt 30 % som rapporteres i Fjeld og Nygaard (2010).

4.1. Evaluering af støj og reduktion af samme

Halskov og Jarvad (2010) evaluerer ikke blot Ordtrawlerens formelle træfrate og genkaldelsesrate, men undersøger også mere kvalitativt hvilke typer støj systemet har svært ved at håndtere og af forskellige årsager fejlagtigt anser for genuine sproglige nydannelser. Formålet med denne "støjanalyse" er at sondere terrænet for mulige forbedringer af systemet (mere herom i afsnit 5).

De primære støjkategorier omfatter

- Bøjningsform af kendt ord
- Banal sammensætning eller lejlighedsdannelse
- Fagsprog
- Stavefejl

De bøjede former af allerede kendte ord er den støjtype som volder Ordtrawleren de største problemer. Årsagen er at ukendte ordformer (altså ordformer som ikke kan henføres til noget lemma i Retskrivningsordbogen 2001) ikke lemmatiseres, og da opslagsordene i eksempelvis Dansk Sprognævns Ordsamling, naturligt nok, står opført i grundformen (se tabel 2), så elimineres de bøjede former

ikke. Den oplagte forbedringsmulighed er naturligvis at anvende et lemmatiseringsprogram som via grammatiske regler kan gætte på grundformen af ord der ikke optræder i dets interne ordbog.

Banale sammensætninger og lejlighedsdannelser er meget svære for Ordtrawleren at skelne fra de nye, blivende ord. Korpuseksempler på førstnævnte er *forskningskvalitet* og *fodboldekspert*, mens *kommunedans* og *pizzabande* er eksempler på de mere uigennemskuelige lejlighedsdannelser. At fjerne disse typer støj maskinelt vil være meget svært. Som beskrevet i afsnit 5 er det eneste håb sandsynligvis at anvende diakron frekvensanalyse.

At skelne fagsprog fra almensprog er, også for mennesker, en ikke-triviel opgave, og for maskinen er det særligt svært da både almensproglige nydannelser og etablerede fagsproglige udtryk ofte vil forekomme nogenlunde lige sjældent i et almensprogligt korpus. Der er dermed ikke noget oplagt statistisk kriterium med hvilket fagsprog kan filtreres fra uden at man også derved frasorterer en række genuine almensproglige nydannelser.

I forhold til støjkategorien stavefejl viste evalueringen at 65 % af disse vedrører brugen af bindestreg (fx *denial-of-service-angreb* og *nul-tolerancepolitik*). Der ligger således en oplagt systemforbedring i at implementere en simpel algoritme som ikke blot sammenholder den eksakte tekststreng for hver nyordskandidat med det allerede kendte ordforråd, men desuden tager samtlige bindestregsvarianter af nyordskandidaten i betragtning.

5. Videreudvikling af Ordtrawleren

I dette afsnit skitseres to oplagte teknikker som kunne tages i anvendelse for henholdsvis at få Ordtrawleren til at identificere mere "avancerede" sproglige nydannelser, såsom ændrede selektionsrestriktioner, og for at eliminere mere vanskelige støjtyper, fx banale sammensætninger og kometord.

5.1. Brug af DanNet til at identificere ændrede selektionsrestriktioner

DanNet-ontologien (Pedersen et al. 2009) er et almensprogligt leksikalsk-semantisk ordnet som er opbygget efter forbillede af Princeton WordNet (Fellbaum 1998), og som siden marts 2009 har været frit tilgængeligt som elektronisk resurse via hjemmesiden http://wordnet.dk.

DanNet består af et antal "synsets" (bundter af synonymer som tilsammen kan anses for at udpege et bestemt begreb), et antal semantiske relationer (fx "used_for" og "has_hyperonym") og endelig et antal ontologiske typer som hvert synset kan siges at repræsentere. Grundlæggende er der to hierarkier af ontologiske typer i DanNet, nemlig

1. konkrete genstande ("1st Order Entities")
2. handlinger, hændelser, egenskaber og abstrakte genstande ("2nd Order entities" og "3rd Order entities")

Ontotypen "Artifact" er en "1st Order Entity" som realiseres i sproget af eksempelvis et ord som *kosmetik*. Ontotypen "Property" er en "2nd Order Entity" (fx *glæde* eller *tørhed*), og "Time" er en "3rd Order Entity". De fleste ontotyper som har fundet konkret anvendelse i DanNet (der er omkring 190 forskellige i version 1.1 af resursen), er imidlertid sammensatte ontotyper, fx er synsettet *kniv* klassificeret som "Instrument+Artifact+Object".

I dette afsnit fokuseres der på disse ontologiske typer og deres potentiale i forhold til at identificere ændrede selektionsrestriktioner for et givet lemma i et analysekorpus kontra et (ældre) referencekorpus.

Den grundlæggende hypotese er at man ved at analysere hvilke ontologiske typer et givet lemmas konkrete syntaktiske argumenter repræsenterer på forskellige tidspunkter i et diakront korpus,

kan få et indtryk af om der sker signifikante forskydninger hen mod eller væk fra bestemte grupper af ontotyper. Er dette tilfældet, så repræsenterer ændringerne, antageligvis, ændringer i lemmaets selektionsrestriktioner.

Analysen er i første omgang meget simplistisk og indebærer simpelthen automatiske opslag i DanNet af samtlige lemmaer der optræder i position +1 i forhold til inputlemmaet (en approksimation af den prototypiske objektsposition), og efterfølgende optælling af hvilke, om nogen, ontologiske typer disse lemmaer kan grupperes under. I nærværende artikel skitseres blot et enkelt eksempel på resultatet af denne simplistiske analyse for inputlemmaet *rejse*. Som det ses i tabel 5, har *rejse* tilsyneladende ikke undergået signifikante forandringer i forhold til de selektionsrestriktioner det lægger på sit direkte objekt. Ændringerne er i hvert fald ikke synlige fra perioden 2004-2005 til perioden 2008-2009.

2004-2005			2008-2009		
Rang	**Ontotype**	**Eksempler**	**Rang**	**Ontotype**	**Eksempler**
1	3rdOrder Entity+ Quantity	krav, kapital, penge	1	3rdOrder Entity+ Quantity	krav, kapital, penge
2	Unbounded Event+ Agentive+ Communication+ Social	tiltale, spørgsmål, debat, diskussion	2	Unbounded Event+ Agentive+ Communication+ Social	tiltale, spørgsmål, debat, diskussion
...	...	...	...	...	...

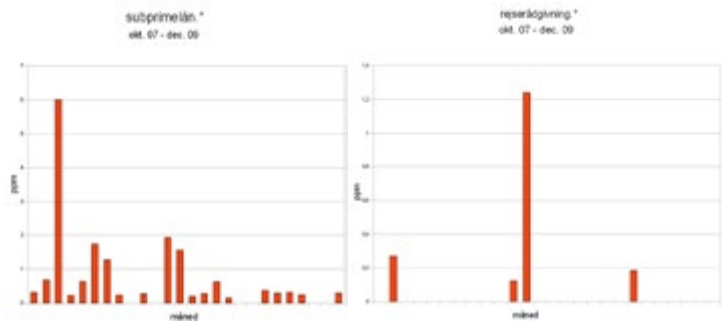
Tabel 5: Brug af DanNet til sammenligning af selektionsrestriktioner over tid: Ontotyper i prototypisk O-position for *rejse*

De to hyppigste ontologiske typer for objektargumenter i 2008-2009 er de samme som i 2004-2005, nemlig abstrakt kvantitet (kapital, penge) og uafgrænsede hændelser af social/kommunikativ karakter (spørgsmål, debat).

Der er naturligvis en række svagheder ved denne teknik. De lemmaer som ikke figurerer i DanNet, bliver eksempelvis ikke henført til nogen ontotype, og desuden er der strengt taget ikke tale om syntaktiske argumenter. Korpus er nemlig ikke er parset, og der anvendes i stedet en simpel positionel SVO-skabelon hvor inputlemmaet er V og lemmaet på position +1 i forhold til V antages at fungere syntaktisk som O. Teknikken skal derfor raffineres således at der som minimum anvendes "phrase chunking". Dermed er det i det mindste kerneleddet i den NP som følger umiddelbart efter V, der slås op i DanNet og ikke det ord som tilfældigvis står i position +1. Under alle omstændigheder virker teknikken anvendelig. Udfordringen består naturligvis i at man ikke på forhånd kan vide hvilke verber der måtte ændre selektionsrestriktioner, og det vil derfor være nødvendigt fuldautomatisk at udtrække og sammenligne data som illustreret i tabel 5 for eksempelvis de 100 eller 1000 mest frekvente danske verber. Denne videreudvikling af Ordtrawleren er på tegnebrættet.

5.2. Inddragelse af diakrone frekvensoplysninger

I dette afsnit skitseres det støjreducerende potentiale i diakrone frekvensprofiler, dvs. afbildninger af den hyppighed hvormed en given nyordskandidat forekommer i et givet korpus som repræsenterer et antal sekventielle tidsperioder (InfoMedia-tekster fra perioden oktober 2007 til december 2009). Figur 2 illustrerer hvordan to nyordskandidater, nemlig *subprimelån* og *rejserådgivning*, har meget forskellige "fingeraftryk" når deres forekomster i et diakront korpus afbildes grafisk.



Figur 2: Diakrone frekvensprofiler

For det første er den relative frekvens af *subprimelån* (0-6 ppm) markant højere end *rejserådgivning* (0-1 ppm). For det andet har *subprimelån* en kometagtig frekvens i begyndelsen af perioden (december 2007) hvorefter frekvensen aftager forholdsvis hurtigt, men dog ikke går i nul. Frekvensprofilen for *rejserådgivning* har et mere cyklisk udseende med nulforekomst i længere perioder (7-9 måneder) afløst af en vis, omend lav, forekomst i enkeltstående måneder.

Hvad kan man så udlede af disse profiler? Sund fornuft fortæller os at *rejserådgivning* er en sæsonbetonet nydannelse, og at *subprimelån* er et kometord som dog næppe forsvinder helt før den nærværende finanskrise er slut. Konklusionen må være at der skal forholdsvis lange tidshorisonter til før diakrone frekvensprofiler kan anvendes til at afsløre kometord. Horisonterne skal sandsynligvis tælles i år snarere end måneder, men selv med sådanne horisonter vil det være vanskeligt at opstille operationelle kriterier som kan danne basis for at en maskine kan afvise, respektive godkende, en nyordskandidats diakrone frekvensprofil.

Omskrives histogrammet for *subprimelån* til en funktion (efter forbillede af Fjeld og Nygaard, 2010), er resultatet tilnærmelsesvist en aftagende eksponentiel funktion, men bør man mekanisk udelukke alle nyordskandidater som følger dette mønster? Histo-

grammet for *rejserådgivning* kunne omskrives til en stykvis funktion som er henholdsvis stigende og aftagende, hvilket måske taler imod udelukkelse.

Selvom kometord kræver en vis tidshorisont at detektere, så vil mange banale sammensætninger givetvis kunne elimineres med diakrone frekvensprofiler hvor diakronien ikke nødvendigvis er omfattende. For eksempel vil sammensætningen *108-116-nederlaget* næppe forekomme i flere forskellige tidsperioder, og selv hvis den gjorde, så ville en minimumsfrekvens på eksempelvis 1 ppm i mindst én periode udelukke den (men hverken *rejserådgivning* eller *subprimelån*).

6. Konklusion og perspektiver

Artiklen illustrerer at det ingenlunde er enkelt at få en maskine til at assistere i arbejdet med at selektere lemmakandidater til en nyordsordbog. Maskinen mangler selvsagt den modersmålskompetence og omfattende omverdensviden og intuition som den menneskelige excerpist er i besiddelse af. Til gengæld kan den excerpere 100 % objektivt og bearbejde store mængder tekst langt hurtigere.

Konklusionen er dog at Ordtrawleren på sit nuværende udviklingsstadium kan udgøre en mærkbar hjælp i forhold til selektionen af de mere simple sproglige nydannelser, dvs. helt nye ord som refererer til nyt indhold. Systemet kan ikke i øjeblikket identificere flerordsforbindelser, nye valensmønstre eller nye selektionsrestriktioner, men som skitseret i afsnit 5, så kan en række sprogteknologiske teknikker og nye resurser tages i anvendelse og sandsynligvis (delvist) udbedre disse mangler.

Litteratur

Ahmad, K. 1993: Pragmatics of specialist terms: The acquisition and representation of terminology. *Machine Translation and the Lexicon, 3rd EAMT Workshop proceedings*.

Drouin, P. 2003: Term Extraction Using Non-technical Corpora as a Point of Leverage. I: *Terminology.* 9(1): 99-117.

Dunning, T. 1993: Accurate methods for the statistics of surprise and coincidence. I: *Computational Linguistics*. 19:1: 61-74.

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

Fellbaum, C. (red.) 1998: *WordNet: An Electronic Lexical Database*. Boston: MIT Press.

Fjeld, R. V. og L. Nygaard 2010: Neologismer i norsk. Kartlegging av leksikalsk språkendring før og nå. I: *Nordiske studier i leksikografi.* Bind 10: 506-521. Rapport fra Konferensen om lexikografi i Norden Tammerfors 3.-5. juni 2009, red. H. Lönnroth og K. Nikula. Tammerfors 2010.

Halskov, J. og P. Jarvad 2010: Human og maskinel excerpering af neologismer. I: *Nydanske Sprogstudier (NyS)* 38, 39-68.

Hansen, D. H. 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.

Heidemann, M. 2009: Om særskrivning, sammenskrivning og brugen af bindestreg i moderne importord. I: *Nyt fra Sprognævnet*. 2009/4: 28-33.

InfoMedia. Danmarks største artikeldatabase. http://www.infomedia.dk.

Jarvad, P. 1995: *Nye ord – hvorfor og hvordan?* København: Gyldendal.

Keson, B. 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.

Pantel, P. og D. Lin 2001: A Statistical Corpus-Based Term Extractor. I: *Lecture Notes in Computer Science.* 2056: 36-46.

Pearson, K. 1904: On the theory of contingency and its relation to association and normal correlation. *Drapers' Company Research Memoirs, Biometric Series 1*. London: Dulau & Co.

Pedersen, B.S., S. Nimb, N., J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen 2009: DanNet: the Challenge of Compiling a Wordnet for Danish by Reusing a Monolingual Dictionary. I: *Language Resources & Evaluation*. 43:3: 269-299.

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

Jakob Halskov
forsker, ph.d.
Dansk Sprognævn
H.C. Andersens Boulevard 2
DK-1553 København V
jhalskov@dsn.dk