NyS

Titel:	Sjældne og sære sammensætninger
Forfatter:	Margrethe Heidemann Andersen og Philip Diderichsen
Kilde:	*NyS – Nydanske Sprogstudier* 41, 2011, s. 40-65
Udgivet af:	NyS i samarbejde med Dansk Sprognævn
URL:	www.nys.dk

NyS 41

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.

Sjældne og sære sammensætninger

- Om særskrivninger og sammensætninger i moderne dansk

MARGRETHE HEIDEMANN ANDERSEN OG PHILIP DIDERICHSEN

Det er en grundregel i dansk grammatik at sammensætninger der udtales med enhedstryk, skrives i ét ord, fx *modebevidst*, *fejlfri* og *tidsgenrer*. Overholdes denne regel ikke, er der tale om særskrivningsfejl som *mode bevidst*, *fejl fri* og *tids genrer*. Flere skandinaviske undersøgelser peger på at særskrivning er en udbredt stavefejltype blandt skoleelever. I og for sig har særskrivning længe været et problem, men antallet af fejl er tilsyneladende stigende, ikke blot i elevtekster (se fx Jervelund 2007), men også i "ellers godt utdannede, voksne og språkbevisste personers skriftspråk" (Vatvedt Fjeld 2004: 14)[1].

Mange har peget på at den øgede tendens til særskrivning i de nordiske sprog kan skyldes indflydelse fra engelsk (se fx Johansson & Graedler 2002: 172; Zola Christensen og Christensen 2005: 39) hvor der ikke som i de nordiske sprog er faste regler for hvordan man skriver sammensætninger, men derimod "stor ortografisk vaklen i brug af sammenskrivning og brug eller ikke-brug af bindestreg" (Steller & Sørensen 1993: 249). Andre forskere sætter dog spørgsmålstegn ved dette udsagn (se fx Walmsness 1999: 121), selvom de fleste vist er enige om at engelsk i det mindste kan have bidraget til udviklingen. Formentlig er der dog også andre faktorer der spiller ind. Vatvedt Fjeld har således peget på at mange særskrivningsfejl undslipper de indbyggede stavekontroller fordi sammensætningens enkelte dele får lov til at passere kontrollen som selvstændige ord (Vatvedt Fjeld 2004: 14), hvilket måske kan være med til at fejltypen udbredes. Dertil kommer at vi med internettets opståen har fået en helt ny situation ift. skriftsproget hvor almindelige mennesker i hidtil ukendt omfang kan læse andre almindelige menneskers tekster i mange til mange-relationer (Duncker 2010: 20) – og dermed også alle de fejl vi måtte lave når vi skriver på nettet.

Selvom der altså er enighed om at særskrivning er et stigende problem, mangler der endnu at blive gjort grundigt rede for særskrivningsfejl i danske (elev)tekster, og der er således tale om et område som mangler at blive belyst i en større, dansk kontekst. Denne artikel tager udgangspunkt i et forskningsprojekt der netop handler om særskrivning og sammenskrivning i moderne dansk, og formålet er at afdække hvilke egenskaber ved sammensætningerne der kan være med til at bevirke at de særskrives. Vores resultater viser hvordan en stor del af særskrivningerne kan undgås hvis man mestrer en håndfuld let identificerbare sammensætningstyper, og resultaterne har dermed også en vis sprogpædagogisk relevans.

HVAD BESTÅR EN SAMMENSÆTNING AF?

Vi undersøger særskrivningsfejl i sammensætninger. Før vi går over til selve beskrivelsen og analysen af særskrivningsfejlene, må vi derfor definere hvad en sammensætning er. Sprogets mindste betydningsbærende enhed er morfemerne, der som bekendt deles op i bøjningselementer, rødder og afledningselementer. Rødder er defineret ved at de kan danne stammer alene, fx *dør* og *greb*. Sættes disse to stammer sammen, danner de sammensætningen *dørgreb*. En sammensætning kan altså defineres som "et ord hvis stamme indeholder mere end én rod" (Hansen & Heltoft 2011: 239). En lignende definition gives i Becker-Christensen & Widell (2000: 133) der skriver at en sammensætning altid "indeholder (…) mindst to rødder". I andre grammatikker defineres sammensætninger som en enhed der består af to selvstændige *ord*, fx "Ordet *husbåt* er satt sammen av de to selvstendige ordene *hus* og *båt*, og vi kaller derfor *husbåt* en **sammensetning**" (Faarlund, Lie & Vannebo 1997: 53). De to forskellige typer definitioner kalder Bondi Johannessen (2001) for hhv. stammeledsanalysen og ordledsanalysen, og hun argumenterer på forskellige måder for at ordledsanalysen ikke er den korrekte måde at analysere en sammensætning på. Hvis en sammensætning består af ord, skriver hun, må man forvente "at både forledd og etterledd kan være bøyd; ubøyde ord (av bøyelig type) kan jo ikke opptre som settningsledd" (2001: 62). Bondi Johannessen fremhæver at bøjning af førsteleddet ikke er muligt hos de to mest

produktive sammensætningstyper, dvs. dem der består af substantiv + substantiv (bortset fra enkelte undtagelser som fx *fædreland*[2]), og dem der består af verbum + substantiv, fx *viskelæder* og *gåstol*. I sammensætninger der består af adjektiv + substantiv, er der eksempler (fx *nytår*, *længstlevende*) hvor førsteleddet tilsyneladende består af et bøjet førsteled. Men, skriver Bondi Johannessen, bøjning er overalt i litteraturen defineret som noget regelmæssigt, og dersom nogle få eksempler som *nytår*, *længstlevende* og *højtflyvende* skulle karakteriseres som bøjning, "ville dette bli en helt ny anvendelse av bøyingsbegrepet, gitt at det mest frekvente åpenbart ikke er bøyde forledd" (2001: 64), jf. sammensætninger som *bredbånd* (ikke *bredtbånd*), *spædbarn* (ikke *spædtbarn*) og *nyord* (ikke *nytord*). Bondi Johannesen argumenterer i stedet for at sådanne sammensætninger består af stammer. En stamme defineres som "den delen av et ord som *ville vært* konstant gjennom hele bøyingsparadigmet *dersom* det forekom som et setningsledd" (2001: 61 f.), hvilket bl.a. er en fordel i forhold til de sammensætninger der indeholder elementer der ikke optræder som selvstændige ord (se nedenfor).

Tyttebær og bomuld

Der findes eksempler på sammensætninger som *tyttebær*, *bomuld* og *stenbider* der indeholder enten forled (hhv. *tytte-* og *bom-*) eller sidsteled (*-bider*) der ikke kan genfindes som selvstændige ord.

Sådanne sammensætninger udgør ifølge Bondi Johannesen (2001: 71) et problem hvis man antager at sammensætninger består af ord (men altså ikke hvis man antager at sammensætninger består af stammer). Og det er hun ikke ene om at mene: Også Aage Hansen har været klar over at en definition af en sammensætning som en forbindelse mellem to eller flere selvstændige ord ikke altid er helt dækkende. Han skriver således:

"Ved adskillelsen af kompositum og komplekst ord plejer man at sige, at det første består af to ord, der forekommer selvstændigt i sproget, medens det sidste indeholder et ord + et element, der ikke forekommer selvstændigt. Men en sådan afgrænsning er ikke helt tilstrækkelig. Efter den vil nemlig fx. cigarmager blive en afledning, fordi –mager ikke forekommer som et selvstændigt ord. En sammenstilling

af cigarmager og fx. cigarfabrikant lader os føle uretfærdigheden i de to tilfælde" (Hansen 1938: 109).

Aage Hansen holder dog i modsætning til Bondi Johannessen fast ved sin definition af sammensæt-ninger som noget der består af ord og konkluderer at der kan være ord "der kun forekommer som led af den særlige art faste forbindelser, vi kalder komposita" (Hansen 1938: 110), ligesom der findes ord der kun forekommer i faste forbindelser som *hip* og *hap* i udtrykket *hip som hap*.

Det vigtige her er ikke hvad de to traditioner kalder sammensætninger som *tyttebær* osv. Pointen er at de begge ganske vist problematiserer dem, men ikke desto mindre kategoriserer dem som sammensætninger, hvilket er argumentet for at vi medtager sådanne sammensætninger i vores analyser. Dertil kommer at der også blandt folk der ikke kender til hverken stammeleds- eller ordledsanalysen formentlig vil være enighed om at *tyttebær*, *stenbider* og *bomuld* er sammensætninger fordi de minder om *jordbær*, *grinebider* og *fåreuld* som alle er uproblematiske i forhold til begge analysemetoder.Vi opererer således med at man kan opdele sammensætninger som *undgå*, *mohair og lørdag* (der alle findes i vores data) i stammerne *und* + *gå*, *mo* + *hair* og *lør* + *dag*, selvom hverken *und*, *mo* eller *lør* findes som selvstændige ord i sproget.

HVILKE VARIABLE ER ASSOCIERET MED SÆRSKRIVNING?

Flere nordiske undersøgelser viser hvilke (og hvor mange) særskrivningsfejl der findes i tekster skrevet af skoleelever og studerende. Disse undersøgelser viser bl.a. at det især er substantivsammensætninger der særskrives, fx *næseoperation* og *vaskebræt* (Walmsness 1999:71; Hallencreutz 2003: 35). Heidemann Andersen har peget på at sammensætninger hvori der indgår proprier (*guccitaske*) eller engelske lån (*tunsteak*), ofte er udsat for særskrivning (Heidemann Andersen 2011), mens Hoaas har påvist at sammensætninger med forkortelser (*DNA-register*), tal (*12-skala*) eller gruppesammensætninger (*cafe- og shoppingmiljøet*) langt oftere særskrives end andre sammensætninger (Hoaas 2008: 59). Hallencreutz (2003) peger på at lange sammensætninger oftere særskrives end korte sammensætninger, og Vollan (2009b:297) har påvist at sam-

mensætninger med en kompleks struktur, uanset ordklasse, særskrives hyppigere end sammensætninger med en enkel struktur. Derudover er det nærliggende at antage at også ordenes frekvens har betydning for om en sammensætning særskrives eller ej. Således foreslår Elbro (2006) at lange, usædvanlige sammensætninger er mere udsatte for særskrivning end andre sammensætninger, dog med den vigtige tilføjelse at det muligvis især er usædvanlige sammensætninger hvis bestanddele er hyppigere end sammensætningen som helhed, hvis bestanddele er genkendelige i udtalen, og hvis bestanddele er genkendelige i betydningen, der særskrives. Elbro giver eksemplet *rødspætte*, der mere eller mindre fejler på alle disse kriterier og derfor formentlig ikke er i risikogruppen for særskrivning. Under alle omstændigheder virker det oplagt at sprogbrugerne lettere husker at sammenskrive højfrekvente ord fordi de simpelthen hyppigt støder på dem i skriftbilledet.

Vi vil i det følgende undersøge en del af de variable der er blevet nævnt som mulige årsager til særskrivningsfejl, nemlig sammensætningernes ordklasse, længde og kompleksitet, om der indgår proprier, engelske lån, forkortelser eller tal i sammensætningerne, og om det har nogen betydning om der er tale om gruppesammensætninger. Vi undersøger også om hyppigheden af såvel sammensætningerne som helhed som hyppigheden af deres bestanddele i forhold til helheden hænger sammen med særskrivningssandsynligheden, samt om der er forskelle i forhold til hvilken type tekst sammensætningerne kommer fra.

DATA OG METODE

Tekster

Det materiale som hidtil er blevet undersøgt med henblik på at finde særskrivningsfejl, udgøres primært af elevtekster, dvs. tekster skrevet af elever på folkeskole- og gymnasieniveau. Vores data, der indeholder i alt 3841 sammensætninger, er bredere sammensat idet særskrivningsfejl jo ikke blot findes i elevtekster, men også i tekster skrevet af voksne, professionelle sprogbrugere. Materialet består derfor af elevtekster, avistekster (herunder annoncer) og blogtekster fra nettet. Elevteksterne består af 17 stile skrevet af htx-elever (Teknisk Skole) på 1. årgang og af 6 bachelorprojekter fra Copenhagen Business School

(CBS). Avisteksterne stammer fra en vestsjællandsk lokalavis (UgeNyt, 5. januar 2011) og er delt op i to kategorier: annoncetekst og redaktionel tekst. Blogteksterne stammer fra websiderne arto.dk, politiken.dk og bt.dk. Den førstnævnte webside er for unge mennesker fra 10 år og opefter. Tabel 1 viser hvor mange løbende ord der ifølge en grov optælling er i hver kilde:

TABEL 1

Kilde	**Antal ord**
Htx-opgaver	15028 ord
CBS-opgaver	64640 ord
UgeNyt-annoncer	1256 ord
UgeNyt-artikler	3721 ord
ARTOblogs	19404 ord
Politikenblogs	11400 ord
BT-blogs	16560 ord
I alt	132009 ord

Variable

Alle sammensætninger[3] i vores data er blevet registreret med angivelse af en række relevante variable[4], nemlig de følgende:

Logfrekvens Dette er et relativt groft mål der skal vise hvor hyppigt en given sammensætning optræder i almensproget: jo højere frekvens en sammensætning har, des mere almindelig er den. Fx har ordet *overførselsråderum* den ekstremt lave værdi -9,1, mens den sammensætning i vores data der har den højeste korpusfrekvens, er *formand* med den højere værdi -3,5. Disse tal er vi kommet frem til på følgende måde: Det blev talt op hvor ofte hver sammensætning forekommer i et stort tekstkorpus bestående af tekster fra Infomedia (en stor samling tekster fra danske aviser, blade og telegrambureauer) med 325 mio. løbende ord, og en rå ordfrekvens blev så foreløbig fundet ved at dividere med 325.000.000. Et ord som *kulturberigere*, der forekommer 3 gange i Infomediakorpusset, får således en rå frekvens på 0.000000009231. Ord

der forekommer 0 gange i dette korpus, blev tildelt fiktive forekomster på 0,5 og 0,25 (hvilket jo egentlig er umuligt – et ord der forekommer, kan ikke forekomme under 1 gang). Dette blev gjort dels for at kunne beregne en logfrekvens for disse ord (se det følgende), dels for at dele ikke-forekommende ord op i to grupper. I gruppen med en fiktiv forekomst på 0,5 findes ord som *minimumstøtte*, *sygehusstrejker* og *ulandsregion* der subjektivt virkede mere almindelige end dem der så blev placeret i den anden gruppe med en fiktiv forekomst på 0,25 som fx *NGO- samt mediehjemmesider*, *børnehavefornærmelse* eller *pro- eller anti-FARC-positioner*. (Se evt. Breland 1996, der også tildeler ikke-forekommende ord fiktive forekomster).

Man må her notere sig en dobbelt skævhed i frekvensmålet. De fiktive forekomsttal er valgt så de giver udtryk for en klart lavere, men ikke ekstremt lav ordfrekvens i et forsøg på at være nogenlunde konservativ og ikke give disse ord en ekstrem overindflydelse i analysen, men i bund og grund er de både subjektivt og arbitrært valgt. Da der er hele 362 ud af de i alt 3841 sammensætninger i vores data der ikke forekommer i Infomediakorpusset, vil de derfor uundgåeligt også tilføre analysen et element af subjektivitet og arbitraritet.

Til gengæld giver de arbitrære, fiktive forekomsttal mulighed for at beregne en rå frekvens på over 0 ved at dividere med 325 mio. (for 0,25 = 0,000000000769, for 0,5 = 0,000000001538), hvilket igen gør det muligt at beregne en logfrekvens som er det endelige frekvensmål der bruges i vores analyse (jf. fx værdien -3,5 for *formand* der teknisk formuleret er en base 10-logaritmetransformation af en rå frekvens på 0,00028; *kulturberiger*es rå frekvens på 0,000000009231 bliver tilsvarende til en logfrekvens på -8,0). Logfrekvens er i hvert fald siden 1960'ernes psykolingvistiske undersøgelser blevet foretrukket frem for rå frekvens som mål for ords hyppighed eller almindelighed, dels fordi ordfrekvenserne bliver mere normalfordelte når de bliver logtransformeret (hvilket er nødvendigt for visse statistiske test), og dels fordi logtransformeret frekvens korrelerer bedre med mange andre sprogrelaterede variable – bl.a. subjektivt vurderet ordhyppighed, jf. note 13 i Tryk (1968): "Since most known correlates of word frequency approximate a logarithmic function of this variable, word frequencies were recorded in log form in this investigation". Se også fx Tanaka-Ishii (2011).

Endelig skal det nævnes som en mulig skævhed i frekvensvariablen at vi har fundet ordfrekvenserne i et korpus som ikke er et rigtigt referencekorpus med et bredt og omhyggeligt balanceret udvalg af tekster i forskellige genrer, som det ellers er idealet i korpuslingvistikken. Vi kunne fx have brugt Korpus 2000 (se Asmussen 2000) der indeholder ca. 28 mio. løbende ord. Problemet her er at hele 752 af vores sammensætninger – 19,6 % – ikke forekommer i dette korpus, hvor tallet for Infomediakorpusset er under halvdelen (362 eller 9,4 %). Vi har ladet korpusstørrelse veje tungere end balanceret sammensætning for at få et frekvensmål for så mange af sammensætningerne som muligt, hvilket i øvrigt turde være forsvarligt da Infomediakorpusset jo trods alt er sammensat af almensprogligt materiale fra aviser og blade.

Del-helheds-frekvens Denne variabel repræsenterer den ene af de mere raffinerede egenskaber ved sammensætninger som Elbro (2006) formoder kan have indflydelse på deres særskrivningssandsynlighed, nemlig delenes frekvens i forhold til helheden. Ud fra den antagelse at sandsynligheden for særskrivning stiger jo mere almindelige de enkelte dele er som selvstændige ord, har vi lavet en talværdi der stiger jo oftere enkeltdelene forekommer i forhold til hele sammensætningen. Dette er gjort ved at dele sammensætningerne op i deres sammensætningsled (af forfatterne som hinandens kontrollanter), optælle enkeltdelenes forekomst i Infomediakorpusset, tage gennemsnittet af enkeltdelenes forekomst, dividere dette gennemsnit med sammensætningens forekomsttal og endelig logtransformere resultatet. Et eksempel på et ord med en høj værdi på denne variabel er *prøvekort* som kun forekommer 1 gang i Infomediakorpusset, mens *prøve* forekommer 25077 gange og *kort* 69379 gange. Gennemsnittet af enkeltdelenes forekomst er 47228, som divideret med *prøvekort*s forekomst på 1 også giver 47228. Logtransformationen af dette giver den endelige værdi for *prøvekort*: 4,7.

Kilde Kilde er en variabel med 7 kategorier. En sammensætning kan således forekomme i en CBS- eller HTX-opgave, en UgeNyt-annonce, en UgeNyt-artikel eller en ARTO-, Politiken- eller BT-blog.

Ordklasse Ordklassevariablen har kategorierne substantiv (fx *tillidshverv)*, verbum (fx *inddraget*), adjektiv (fx *sørgmodige*) og pronomen (fx *ingenting*).

Proprium, Forkortelse, Tal, Engelsk og Gruppesammensætning Alle disse variable har to kategorier. *Proprium* deler data op i de sammensætninger hvor der indgår proprier (fx *guccitaske*), og de sammensætninger hvori der ikke indgår et proprium. *Forkortelse*: Forkortelser indgår i sammensætningerne (fx *DNA-register*) eller ikke. *Tal*: Tal indgår (fx *12-skala*) eller ikke. *Engelsk*: Engelske lån indgår (fx *tunsteaks*) eller ikke. *Gruppesammensætning*: Sammensætningerne er gruppesammensætninger (fx *cafe- og shoppingmiljøet*) eller ikke.

Led Denne variabel er en operationalisering af sammensætningernes kompleksitet og er fundet ved simpelthen at tælle antallet af led. *Stats- og udenrigsministre* har værdien 5 på denne variabel.

Længde Denne variabel er sammensætningernes absolutte længde i antal tegn. Længden er målt på den korrekt sammenskrevne form. Den særskrevne form *latex duo topmadras* i data har derfor en længde på 17.

Overflademønstre i data

Vores datamateriale kan opdeles på kryds og tværs efter disse variable, og alt efter hvilken variabel man tager udgangspunkt i, synes de følgende udsagn at gælde. Enkeltvis er disse udsagn i høj grad i overensstemmelse med tidligere fund: 1) Jo mere almindelige (hyppige) sammensætninger er i sproget, des færre særskrivninger observeres. Der er altså en negativ sammenhæng mellem frekvens og særskrivning. 2) Lidt overraskende *falder* sandsynligheden for særskrivning også når sammensætningsleddenes hyppighed i forhold til hele sammensætningens hyppighed stiger. 3) Der er forskel på særskrivningssandsynligheden afhængigt af hvilken kilde data kommer fra, med færrest særskrivninger i redaktionel tekst fra UgeNyt og flest i htx-stile. 4) Sammensætninger med proprier særskrives langt oftere end sammensætninger uden proprier. 5) Sammensætninger med forkortelser særskrives langt oftere end dem uden. 6) Sammensætninger med tal særskrives oftere

end dem uden. 7) Der er flest særskrivninger blandt sammensætninger som er substantiver, dvs. har substantiv som sidsteled. 8) Der forekommer flere særskrivninger jo flere led sammensætningerne består af. 9) Der forekommer flere særskrivninger jo længere sammensætningerne er. 10) Sammensætninger med engelske lån særskrives oftere end dem uden. 11) Gruppesammensætninger særskrives oftere end ikke-gruppesammensætninger.

Hvis man testede disse mønstre statistisk enkeltvis, fx med χ^2-test, ville man finde at tilsyneladende alle variablene lige fra korpusfrekvens til ordlængde havde en statistisk signifikant sammenhæng med forekomsten af særskrivninger. En sådan tolkning kan imidlertid kun lade sig gøre hvis alle kategorierne er separate, dvs. hvis der fx ikke var nogen proprier i de særskrivninger der indeholder engelske lån, eller hvis der ikke var nogen forkortelser i gruppesammensætningerne. Da det ikke er tilfældet – der er altså overlap mellem kategorierne – bør man stille sig selv spørgsmålet hvilke af variablene der er de vigtigste, og hvor meget resten betyder når der er taget højde for dem. Det kan man gøre ved hjælp af en multipel regressionsanalyse, i dette tilfælde logistisk regression, som vi i det følgende afsnit vil præsentere.

Analysemetode

Multipel logistisk regressionsanalyse kan måle sammenhængen mellem flere variable og et binært udfald (her: særskrivning eller ikke). Analysen beregner en ligning for sammenhængen mellem alle variablene og udfaldet. Ligningen, kaldet en statistisk model, og mere specifikt en logistisk regressionsmodel, indeholder en række koefficienter, dvs. faktorer der vægter de forskellige variable. Under beregningen af modellen vægtes en variabel positivt hvis sandsynligheden for særskrivning stiger med variabelværdien, men negativt hvis sandsynligheden falder. Fx vil længde blive vægtet positivt fordi man kan sige at jo længere et ord er, des højere er sandsynligheden for at det bliver særskrevet. Omvendt falder sandsynligheden for særskrivning jo mere hyppigt et ord er, så frekvens vil blive vægtet negativt. Et ord kan sættes ind i en færdig model i form af en række variabelværdier, fx: (1; 0; 0; 0; 1; 24; 4; -9,1) for ordklasse = substantiv; proprium = ja; engelsk lån = nej; forkortelse = nej; tal = nej; gruppesammensætning = ja; længde = 24; antal led = 4;

logfrekvens = -9,1 – svarende til den observerede ordform *Nørgaard på strøget pose*.[5] Når ordet sættes ind i modellen, vil værdierne blive ganget med vægtene (dvs. de vægtede koefficienter), lagt sammen og resultere i en sandsynlighed for at ordet er særskrevet.[6] Målet med modelberegningen er at indrette modellen således at den for hvert ord giver den mindst mulige forskel på den beregnede særskrivningssandsynlighed og den observerede sandsynlighed (her = 1 fordi gruppesammensætningen *Nørgaard på strøget pose* kategoriseres som særskrevet eftersom den mangler en bindestreg). En anden indfaldsvinkel til denne pointe er at beregningen skal resultere i en model der så vidt muligt tildeler en særskrivningssandsynlighed på over 0,5 (svarende til en kategorisering som særskrevet) til ord der faktisk er særskrevet, og en sandsynlighed på under 0,5 til ord der ikke er særskrevet. Dette gør den logistiske regressionsanalyse ved at finde frem til de modelkoefficienter der bedst opfylder disse mål.

Det er ikke sikkert at alle de forhåndenværende variable kan bidrage til den bedst mulige model, og som en vigtig del af analysen kasseres de variable der ikke bidrager.[7] Når en optimal model er fundet med et bestemt udvalg af variable, vil det fremgå hvilke variable der har den største betydning i det samlede billede. Det springende punkt er at en given model i analysen af data tager højde for alle de andre variable når den beregner betydningen af en given variabel. Således er det ikke nødvendigvis alle variable der har en selvstændig sammenhæng med særskrivning; kun dem der kan forbedre modellens estimater af de enkelte ords særskrivningssandsynlighed ud over det der er opnået af de andre variable tilsammen, kan tillægges selvstændig betydning.

Ud over at den kan tage højde for flere variable samtidig, er det en vigtig egenskab ved logistisk regressionsanalyse at den i modsætning til en simpel χ^2-analyse kan operere med kontinuerte variable, fx den reelle længde af hvert ord målt i antal bogstaver, samtidig med at den håndterer det binære udfald (særskrivning eller ikke). Den logistiske regression tager således alle ordene med fx 10 bogstaver og beregner en sandsynlighed for fejl blandt disse. Det samme sker for ord med alle andre længder i data. Derefter finder regressionsanalysen den modelkoefficient der bedst opsummerer alle disse sandsynligheder, og hvis der er en tilstrækkeligt stigende eller faldende tendens i forhold til antallet

af bogstaver, vil ordlængdens sammenhæng med fejlsandsynligheden være statistisk signifikant: En lav p-værdi (< 0,05) udtrykker i så fald at stigningen eller faldet i sandsynlighed er for stort til at være tilfældigt.

I det følgende beskriver vi resultaterne af en logistisk regressionsanalyse af alle vores variable.

RESULTATER

Den statistiske model for særskrivningssandsynligheden i vores data er opbygget ved hjælp af en automatisk trinvis procedure (se Johnson 2008). Proceduren tilføjede variable en efter en ud fra hvilken variabel der på et givet trin kunne bidrage mest til modellen. Den model der var resultatet af denne manøvre, indeholdt de variable der fremgår af statistikprogrammet R's output nedenfor, der altså udgør analysens resultat: logfrekvens, kilde, proprium, forkortelse, ordklasse, led og tal.

```
> anova(glm(formula = fejl ~ logfrekv + kilde.un + pro-
prium + fork + ordklasse + led + tal, family = binomial,
data=s), test="Chisq")

Analysis of Deviance Table
Model: binomial, link: logit
Response: fejl
Terms added sequentially (first to last)

          Df  Deviance  Resid.Df Resid. Dev P(>|Chi|)
NULL                      3840     1854.7
logfrekv. 1     500.74    3839     1354.0 < 2.2e-16 ***
kilde     6      75.81    3833     1278.2 2.619e-14 ***
proprium  1      43.24    3832     1234.9 4.838e-11 ***
fork.     1      39.32    3831     1195.6 3.593e-10 ***
ordklasse 3      22.10    3828     1173.5 6.223e-05 ***
led       1      16.03    3827     1157.5 6.241e-05 ***
tal       1      11.00    3826     1146.5  0.000911 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1' '1
```

Outputtet viser de enkelte variables bidrag til den færdige model når der tages højde for de variable der står over en given variabel i outputtet. Alle variablene er signifikante og viser dermed en signifikant sammenhæng med særskrivning ud over de foregående variable.

Outputtet skal forstås som følger. Det centrale i outputtet er kolonnen "Resid. Dev" (*residual deviance* 'resterende afvigelse'). Den viser et mål for hvor meget modellen på et givet trin, dvs. med et givet antal variable, stadig afviger fra en perfekt model. En perfekt model giver hver særskrivning i data en særskrivningssandsynlighed på 1 og hver sammenskrivning en særskrivningssandsynlighed på 0 – dvs. den "gætter rigtigt" hver gang, og der er ingen afvigelse mellem den tilskrevne sandsynlighed for særskrivning (hhv. 1 og 0) og den faktiske sandsynlighed (også hhv. 1 og 0). En uperfekt model giver sammensætningerne en sandsynlighed på et sted mellem 0 og 1 for at blive særskrevet, og de kommer således til at afvige fra den perfekte models særskrivningssandsynligheder. "Resid. Dev" er et totalmål for disse afvigelser: jo flere og større afvigelser, des højere "Resid. Dev". Værdien i den øverste linje ud for "NULL" er udtryk for en helt tom models samlede afvigelse fra det perfekte. En tom model deler ikke data op efter nogen variable og tilskriver derfor per default alle sammensætninger i data den samme sandsynlighed for at blive særskrevet. Det svarer til at sige: "Der er 6,53 % særskrivninger i data. Derfor går vi ud fra at en sammensætning har 6,53 % sandsynlighed for at blive særskrevet uanset hvor sjælden den er, og uanset hvad der indgår af usædvanlige orddele såsom proprier og forkortelser osv.". Altså at fx *3-d-CT-skanninger* (det er den korrekte stavemåde ifølge Retskrivningsordbogen!) har samme sandsynlighed for at blive særskrevet som *oplyser*, for nu at tage et par ekstremer fra vores data. Det kan selvfølgelig differentieres bedre, og det er det der sker efterhånden som de enkelte variable tilføjes. Med tilføjelsen af variable forbedres modellen idet afvigelsen fra en perfekt model ("Resid. Dev.") falder. Men det man også skal lægge mærke til, er hvordan hver tilføjet variabel tenderer til at være mindre vigtig end den foregående. Forbedringen af modellen aftager gradvist efterhånden som man bevæger sig ned gennem tabellen. Det er udtrykt direkte i "Deviance" i ouputtet, som angiver hvor meget modellen forbedres variabel for variabel – og det er altså mindre og mindre. Effekten af

hver tilføjet variabel er dog statistisk signifikant, hvilket man kan se på de meget små p-værdier i kolonnen "P(>|Chi|)", der kommer fra en χ^2-test af "Deviance"-værdien. I det følgende kommenterer vi hver af variablene for sig.

Frekvens

Af de variable vi har undersøgt, er det, som det fremgår af outputtet ovenfor, logfrekvens der er den vigtigste. Jo mere sjældent et sammensat ord er, des højere er sandsynligheden for at det bliver særskrevet – eller måske snarere omvendt: Jo mere almindeligt (dvs. hyppigt) et sammensat ord er, des lavere er særskrivningssandsynligheden.

Vi har også undersøgt om forholdet mellem hele sammensætningens frekvens og delenes frekvens har betydning for særskrivningssandsynligheden (jf. Elbro 2006). Del-helheds-frekvens kom aldrig med i modellen. Vores analyse peger altså ikke på at der er større sandsynlighed for særskrivning hvis delene i en sammensætning er mere frekvente end sammensætningen som helhed – givet de andre variable i modellen. Ganske vist er delene i de særskrevne ord i gennemsnit hyppigere end helheden – det gælder imidlertid også for de sammenskrevne ord, men bare i endnu højere grad: Gennemsnittet for variablen er 1,18 for de særskrevne ord og 1,84 for de sammenskrevne, og forskellen er signifikant (Mann-Whitney $U = 330534{,}5$; $p < 0{,}00001$). Det viser sig at førsteleddet i sammenskrevne sammensætninger er mere end en størrelsesorden hyppigere end førsteleddet i særskrivninger med gennemsnitlige logfrekvenser på hhv. -4,0 og -5,3 ($W = 680555{,}5$; $p < 0{,}00001$), og det er formentlig en stor del af forklaringen. Som vi kommer tilbage til nedenfor, indeholder særskrivninger fx ganske ofte proprier eller forkortelser, og det virker rimeligt at varenavne som *Tempur* i *Tempurmaterialet* eller *Dunlopillo* i *Dunlopillokup* og forkortelser som *ECTS*, *STX*, *SOSU* og *NGO* er sjældne i almensproget.

Kilde

Den næstvigtigste variabel er kilde, det vil sige at det gør en forskel hvilken type tekst sammensætningerne forekommer i. Fordelingen af datamaterialet når det stilles op alene efter tekstkilde, giver et fingerpeg om hvorfor denne variabel er betydningsfuld.

TABEL 2

Kilde	Sammenskrevet	Særskrevet	Total	Procent særskrivninger
Annonce	351	53	404	13,1
ARTO	259	40	299	13,4
BT	375	23	398	5,8
CBS	1717	60	1777	3,4
Politiken	369	25	394	6,3
Htx	260	43	303	14,2
UgeNyt	259	7	266	2,6
Total	3590	251	3841	6,5

Det fremgår således af ovenstående tabel at der er relativt mange særskrivninger i UgeNyts annoncer. Det stemmer fint overens med at undersøgelser har vist at der er mange særskrivninger i tekster med et særligt visuelt design eller layout, fx annoncer (Mobärg 1997). Til sammenligning er det redaktionelle stof i den samme lokalavis den teksttype der indeholder færrest særskrivninger, hvilket næppe kan undre i betragtning af at disse tekster er skrevet af professionelle skribenter – og at teksterne derudover bliver korrekturlæst. UgeNyts annoncer er kun overgået af ARTO-brugere og htx-elever, mens CBS-studerende har relativt få særskrivninger. Politiken- og BT-bloggere har en lavere særskrivningssandsynlighed end htx-elever og ARTO-brugere. Der tegner sig et billede af tre forskellige grupper der ikke nødvendigvis er særlig homogene. UgeNyt-annoncer, ARTO-brugere og htx-stile indeholder mange særskrivninger, Politiken- og BT-blog-tekster noget færre og CBS-opgaver og redaktionel tekst fra UgeNyt færrest.

Proprier, forkortelser og tal

De næste variable på listen er proprium og forkortelser. Sammensætninger hvori der indgår et proprium (fx *guccitaske*) eller en forkortelse (fx *DNA-register*), er således mere tilbøjelige til at blive fejlstavet end sammensætninger hvor disse elementer ikke indgår. Det samme gælder tal (fx *12-skala*), selv om variablen tal har mindre betydning end de to andre.

Ordklasse

Også variablen ordklasse spiller en rolle for særskrivningssandsynligheden. Tidligere undersøgelser viser som nævnt at det især er sammensætninger med substantiver som sidsteled der særskrives (fx *næseoperation* og *danskstuderende*), og dette mønster går igen i vores data, som man kan se i nedenstående tabel.

TABEL 3

Ordklasse	Sammenskrevet	Særskrevet	Total	Procent særskrivninger
Adjektiv	390	20	410	4,9
Pronomen	9	0	9	0,0
Substantiv	2397	231	2628	8,8
Verbum	794	0	794	0,0
Total	3590	251	3841	6,5

Her bemærker man at de to af kategorierne slet ikke indeholder nogen særskrivninger. Der findes altså ingen særskrevne verber eller pronomener i vores data. For at være sikker på at ordklassevariablen ikke havde nogen forvrængende effekt på modellen, hvilket kan forekomme ved kategorifrekvenser på 0 på grund af den logistiske regressionsanalyses beskaffenhed (se også note vi), fjernede vi sammensætninger der var verber og pronomener fra datasættet og kørte modellen igen. Outputtet ses her:

```
> anova(glm(formula = fejl ~ logfrekv + kilde.un + pro-
prium + fork + ordklasse + led + tal, family = binomial,
data=suv), test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: fejl
Terms added sequentially (first to last)
```

```
          Df  Deviance  Resid.Df Resid. Dev P(>|Chi|)
NULL                       3037     1732.4
logfrekv. 1     405.18     3036     1327.2 < 2.2e-16 ***
kilde     6      72.22     3030     1255.0 1.433e-13 ***
proprium  1      43.25     3029     1211.8 4.807e-11 ***
fork.     1      37.97     3028     1173.8 7.186e-10 ***
ordklasse 1       0.28     3027     1173.5  0.593538
led       1      16.03     3026     1157.5 6.241e-05 ***
tal       1      11.00     3025     1146.5  0.000911 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1' '1
```

Det reducerede datasæt resulterer ikke i de store forandringer, bortset fra at ordklassevariablens bidrag til modellen nu ikke længere er signifikant. Det er dels et tegn på at variablene faktisk står i den rigtige rækkefølge efter vigtighed, og dels at der ikke er nogen signifikant forskel på substantiver og adjektiver i vores data hvad særskrivningssandsynlighed angår. Det er altså først og fremmest verberne der skiller sig ud ved ikke at blive særskrevet.

Længde og antal led

Endelig spiller antallet af led i sammensætningerne en rolle: Jo flere led en sammensætning består af, des større er sandsynligheden for at den bliver særskrevet. Derimod har sammensætningens længde ifølge vores analyse ingen selvstændig sammenhæng med særskrivningssandsynligheden. Det skyldes at den oversandsynlighed for særskrivning der isoleret set er forbundet med længde, kan forklares med andre variable. Således korrelerer især logfrekvens negativt med længde, dvs. at lange ord tenderer til at være lavfrekvente (r = -0,57). Og fordi frekvens så suverænt forklarer en stor del af sandsynligheden for særskrivning, er der så at sige ikke meget tilbage at forklare for variablen længde. Antal led korrelerer ikke i lige så høj grad med frekvens (r = -0,33). Sammensætninger med mange led tenderer altså i lidt mindre grad til at være lavfrekvente. Der er således plads til at lidt mere højfrekvente ord der

ikke er lange, men som har relativt mange led, kan forklare en smule af særskrivningssandsynligheden.

Engelsk og gruppesammensætning

Engelske lån (fx *timeout* og -*steak* i *tunsteak*) og gruppesammensætninger (fx *cafe- og shoppingmiljøet*) spiller heller ingen selvstændig rolle for særskrivning i denne analyse.[8] Det skyldes at de simpelthen overskygges af logfrekvensvariablen. Sammensætningerne med engelske lån er generelt sjældne, med undtagelse af ordet *weekend*. For gruppesammensætningerne gælder det i ekstrem grad: Ingen af de 53 gruppesammensætninger i vores data forekommer en eneste gang i Infomediakorpusset. De to sammensætningstyper er simpelthen sjældne, og derfor har de to variable ikke noget yderligere at bidrage med når først det er sagt at lavfrekvente sammensætninger tenderer til at blive særskrevet.

Sprogpædagogisk relevante sammensætningstyper

Ligesom man kan tage udgangspunkt i sammensætninger i almindelighed for at undersøge de variable der hænger sammen med særskrivning, som vi har gjort i de foregående afsnit, kan det have sin egen relevans at se på hvad man kan få ud af at eliminere forskellige let identificerbare typer af særskrivninger. Der er en række velafgrænsede sammensætningstyper der ofte særskrives: Sammensætninger hvor der indgår proprier eller forkortelser (hhv. 70,9 % og 65,4 % særskrivninger), gruppesammensætninger (49,1 % særskrivninger) og sammensætninger hvor der indgår engelske lån eller tal (hhv. 32,1 % og 27,8 % særskrivninger). Selv om hver af disse kategorier forekommer relativt sjældent, er de så udsat for særskrivning at de hver især udgør en pæn andel af de 251 særskrivninger i vores data, se tabel 4:

TABEL 4

Sammensætningstype	Andel sammensætninger af denne type blandt de i alt 3841 sammensætninger	Andel særskrivninger af denne type blandt de 251 særskrivninger
Proprium	2,2 %	15,5 %
Forkortelse	1,4 %	13,5 %
Engelsk	1,4 %	10,8 %
Gruppesammensætning	0,9 %	10,4 %
Tal	1,4 %	4,0 %

Desuden er overlappet mellem disse kategorier så lille, at hvis man helt eliminerede særskrivninger af disse 5 typer i vores data, ville man fjerne 101 af de 251 særskrivninger (40,2 %). Engelsk og gruppesammensætning har ifølge vores analyse ingen selvstændig sammenhæng med særskrivning. Således er det ud fra vores resultater fx ikke de engelske lån *som sådan* der gør det svært at skrive sammensætninger korrekt, men snarere sammensætningers usædvanlighed (som engelske lån på den anden side godt kan tænkes at bidrage til). Men det kan være relevant for sprogpædagogikken at fokusere på reglerne for disse 5 sammensætningstyper fordi de er så nemme at identificere.

DET SJÆLDNE OG DET SÆRE

Formålet med denne artikel har været at undersøge vigtigheden af en række egenskaber ved sammensætninger der hænger sammen med særskrivning. Hovedresultatet af vores undersøgelse er at en given sammensætnings logfrekvens er den enkeltvariabel der bedst kan forudsige om den særskrives eller ej. Jo mere usædvanlig en sammensætning er, des højere er sandsynligheden for at den bliver særskrevet. Dette resultat støttes af Vatvedt Fjeld som i et pilotprojekt om særskrivning og sammenskrivning i norsk finder at "sammensetninger som er forholdsvis nye eller sjeldne", kun forekommer i særskrevet form (Vatvedt Fjeld 2004: 21). Det tyder på at der er en vis sammenhæng mellem særskrivning og genkendelighed; sammensætninger der er gamle, veletablerede og frekvente særskrives sjældnere end sammensætninger

der er nye, uetablerede og lavfrekvente. Man kan spørge om sammensætninger er svære at skrive korrekt fordi de er sjældne – eller om de snarere er sjældne fordi de er nye og kreative. Med andre ord: Kan sjældenhed som sådan være en årsag til særskrivning? Svaret giver til en vis grad sig selv. Det er svært at forestille sig at ords sjældenhed i sig selv skulle drive folk til at særskrive dem: Jo mindre man har set de sjældne former, des mindre må det antages at de kan påvirke en til at særskrive dem. Så er det mere rimeligt at foreslå at hyppighed kan drive korrekt sammenskrivning. Så hvorfor særskrives sjældne sammensætninger? Måske laver folk simpelthen et mellemrum hver gang de mener at have skrevet et helt ord, når den sammensætning de er i gang med at skrive, ikke er en de har erfaring med, som foreslået i Elbro (2006). Vores data giver dog ikke mulighed for at efterprøve denne eventuelle tilbøjelighed hos de enkelte skribenter.

De enkelte skribenter dukker til gengæld op i forbindelse med den næstvigtigste variabel vi har identificeret, nemlig hvilken tekstype sammensætningen kommer fra, dvs. hvilken type skribent der er ophavsmand til sammensætningen. Her viser vores resultater at det er de formentlig yngste skribenter, dvs. ARTO-skribenterne og htx-skribenterne, der laver flest særskrivningsfejl, mens de skribenter der laver færrest særskrivningsfejl, er ældre og enten er i gang med en længerevarende uddannelse (CBS-skribenterne) eller er professionelle skribenter (journalisterne på lokalavisen). Da vi ikke har aldersinformation på skribenterne i vores data, har vi dog ikke mulighed for at sige noget om hvorvidt unge skribenter generelt laver flere særskrivningsfejl end ældre.

Ud over korpusfrekvens og tekstkilde har en række andre variable også betydning, dog af aftagende vigtighed. Det har betydning om der indgår et proprium i sammensætningen – måske fordi man har lyst til grafisk at markere propriet som noget selvstændigt. Det har også betydning om der indgår en forkortelse eller et tal i sammensætningen. Det peger bl.a. på at Retskrivningsordbogens § 63 om brugen af bindestreg i sammensætninger med tal og forkortelser ikke kendes af en stor del af sprogbrugerne. Også gruppesammensætninger falder ind under Retskrivningsordbogens § 63, og som det fremgår af tabel 4, er der en høj andel af særskrivningerne der findes i de ovenstående

kategorier. Der kan således være behov for flere undersøgelser for at klarlægge i hvor høj grad manglende viden om § 63 i Retskrivningsordbogen er årsag til særskrivningsfejl.

Afslutningsvis vil vi pege på at der uden tvivl mangler vigtige variable før vi kommer i nærheden af en model der begynder at give et dækkende billede af hvad der bestemmer særskrivning. Der er flere interessante forslag i Elbro (2006), ikke mindst ville vi gerne have kunnet inddrage flere egenskaber ved de skribenter der producerer de sammensatte ord. En psyko- eller sociolingvistisk tilgang ville være interessant, men er selvsagt uden for rækkevidde her – hvad angår skribentegenskaber, har vi kun den grove kategorisering efter tekstkilder at gøre godt med.

Det vi har leveret i denne artikel, er således en rangordning af et udvalg af variable der har været tilgængelige – først og fremmest ordvariable. Resultaterne viser at det er de sjældne og – meget passende – de sære sammensætninger der ofte særskrives.

Margrethe Heidemann Andersen
Dansk Sprognævn
heidemann@dsn.dk

Philip Diderichsen
Dansk Sprognævn
phildi@dsn.dk

NOTER

1 Også i det materiale der vil blive analyseret i denne artikel, er særskrivninger et relativt alvorligt staveproblem – det er lige så svært for staverne at undgå særskrivningsfejl som det er at undgå *r*-fejl, dvs. den notoriske fejlstavning af de enslydende endelser *-er*, *-re* og *-rer* (se Jervelund & Schack 2010: 53). Andelen af fejl blandt verber med disse endelser og andelen af særskrivninger blandt sammensætninger i vores data er således omtrent den samme (hhv. 5,2 % og 6,5 % – forskellen er ikke signifikant, $\chi^2 = 1{,}63$; df = 1; p = 0,2). (Da de fleste data ikke har foreligget i elektronisk form, har vi været nødsaget til at begrænse sammenligningsgrundlaget til verber. Vi antager dog at fejlprocenten blandt verber med r-fejl-basis er nogenlunde den samme som for andre ordklasser).

2 Det skal understreges at der ikke behøver være talkongruens mellem forleddet og efterleddet, jf. eksempler som *småpike* og *fædreland* hvor der ifølge Bondi Johannesen "godt kan være bare en far med i bildet (mitt fedreland)" (Bondi Johannesen 2001: 64).

3 Vi har dog udeladt alle adverbier, præpositioner, konjunktioner og udråb, vel vidende at mange særskrivningsfejl netop findes blandt disse ordklasser (se fx Jervelund 2007). Når vi har valgt at udelade disse ordklasser fra vores undersøgelse, skyldes det at vi har valgt at fokusere på de særskrivningsfejl der bryder med de traditionelle grammatikregler på området (fx reglen om hovedledstryk, se Heidemann Andersen (2011)), snarere end særskrivningsfejl der i højere grad er brud på mere arbitrære regler (jf. Hansen 1967: 211). Selvom der i og for sig er tale om brud på retskrivningsreglerne i begge tilfælde, er der enighed om at særskrivningsfejl ved adverbium + præposition er mindre alvorlige (jf. fx forordet Retskrivningsordbogen 1955), og reglerne for sammenskrivning og særskrivning af adverbium + præposition bliver da også lempet i næste udgave af Retskrivningsordbogen, der forventes at udkomme i 2012 (se Schack 2011). Vi har derfor vurderet at det i en undersøgelse af særskrivningsfejl i moderne dansk ikke vil være hensigtsmæssigt at inddrage de særskrivninger som er fejl i dag, men ikke vil være det efter 2012. Havde vi medregnet alle typer særskrivningsfejl, ville de tal vi opererer med, være uaktuelle efter 2012, og det ville efter vores mening være uheldigt. Dertil kommer at sprogbrugerne oftere laver sammenskrivningsfejl end særskrivningsfejl i disse forbindelser (*han er født indenfor voldene* i stedet for *han er født inden for voldene,* jf. Schack 2011: 3), hvilket er yderligere et argument for ikke at medtage disse ordklasser i en undersøgelse hvor særskrivningsfejl er i fokus.

4 Tak til stud.mag. Sofie Vestergaard Bräuner for hjælp med excerperingen og indtastningen af materialet.

5 NB: Et par variable er udeladt i dette eksempel.

6 Det skal for god ordens skyld tilføjes at logistisk regression ikke opererer på rå sandsynligheder, men derimod på en logaritmetransformation af sandsynlighederne hvor en sandsynlighed på 0 % svarer til -∞ (minus uendelig), 50 % svarer til 0, og 100 % svarer til +∞. Det er bl.a. for at undgå sandsynligheder uden for intervallet 0-100 %, hvilket ville være meningsløst. (En fejl kan fx ikke have mindre end 0 % eller mere end 100 % sandsynlighed for at forekomme, uanset hvor kort eller lang en sammensætning der så i øvrigt er tale om).

7 Det kan gøres på flere måder, fx ved at starte med en model der indeholder samtlige variable og så en efter en fjerne dem der ikke har nogen statistisk signifikant betydning for særskrivning, eller man kan bygge en model op ved at tilføje de variable der har størst betydning givet effekten af de foregående en efter en. Man kan også på forhånd udvælge nogle variable som man er interesseret i hvor langt man kan komme med uden at gå så meget op i hvor stor en effekt de egentlig har i forhold til andre variable. Denne tilgang er dog helt uaktuel her, hvor hele formålet er at finde frem til de variable der har størst effekt uden specifikke a priori-antagelser.

8 Engelske lån dukker op i modellen længere nede i den automatiske trinvise procedure, men først efter et punkt hvor modellen ikke længere har nogen rimelig generaliserbarhed. Den trinvise procedure tilføjede adskilligt flere variable end der er med i modellen ovenfor, men mange af disse måtte fjernes igen af hensyn til generaliserbarheden. Alt andet lige er det hensigtsmæssigt at have få parametre i en statistisk model hvis den skal være en god beskrivelse også af data som ikke er brugt til at beregne modellen. Det illustreres måske bedst af det ekstreme tilfælde hvor man har en model med lige så mange parametre som der er sammensætninger. I så fald vil en model kunne give den nøjagtige sandsynlighed for hvert ord (dvs. 1 eller 0) og på den måde ramme rigtigt hver gang – men vel at mærke kun med de ord der optræder i data. Modellen vil nemlig blive så idiosynkratisk at tilskrivningen af særskrivningssandsynligheder til nye data bliver kaotisk og helt hen i vejret. Det kan allerede have katastrofale konsekvenser for generaliserbarheden hvis en model har nogle få variable for meget. Derfor blev de mindst vigtige parametre fjernet indtil modellen når den blev beregnet på kun 85 % af data, i gennemsnit kunne give en rimelig forudsigelsesprocent (omkring 94,5 % i gennemsnit) på de sidste 15 % af data der ikke var brugt i beregningen af modellen. (Se Johnson 2008).

LITTERATUR

Asmussen, J. (2002) Korpus 2000. "Et overblik over projektets baggrund, fremgangsmåder og perspektiver". I: *Nys30-Korpuslingvistik*. København: Akademisk Forlag.

Bauer, L. (1978) *The Grammar of Nominal Compounding with special reference to Danish English and French*. Odense: Odense University Press.

Bauer, L. (1983) *English Word-formation*. Cambridge: Cambridge University Press.

Becker-Christensen, C. & P. Widell (2000) *Politikens Nudansk Grammatik*. København: Politikens Forlag.

Bondi Johannessen, J. & H. Hauglin (1998) An automatic analysis of Norwegian compounds. T. Haukioja (red.): *Papers from the 16th Scandinavian Conference of Linguistics*, Turko/Åbo, Finland 1996. Turko: University of Turko, Department of Finnish and General Linguistics. 209-220.

Bondi Johannessen, J. (2001) "Sammensatte ord". *Norsk Lingvistisk Tidsskrift*. Årgang 19. 59-91.

Duncker, D. (2010) "Dansk sprognormering nu og i fremtiden". *Sprog i Norden 2010*. Netværket for sprognævnene i Norden. Nordisk Ministerråd. 19-31.

Breland, H. M. (1996) *Word frequency and word difficulty: A comparison of counts in four corpora. Psychological Science*. 96-99. JSTOR.

Elbro, C. (2006) "Sammensatte ords deling. Sær skrivning". *Mål og Mæle* 1/2006. 12-19.

Faarlund, J. T., S. Lie & K. I. Vannebo (1997) *Norsk Referansegrammatikk*. Oslo: Universitetsforlaget.

Hallencreutz, K. (2001) "Skyll på längden, intet på engelskan". *Språkvård* 2001/4. 4-9.

Hallencreutz, K. (2003) *Särskrivningar och andra skrivningar i elevspråk*. Uppsala: Uppsala universitet, FUMS.

Hansen, A. (1938) *Indledning til nydansk Grammatik*. Århus: Universitetsforlaget i Århus

Hansen, A. (1967) *Moderne dansk*, bind I-III. København: Grafisk Forlag.

Hansen, E. & L. Heltoft (2011) *Grammatik over det Danske Sprog. Bind 1. Indledning og oversigt*. Det Danske Sprog- og Litteraturselskab.

Heidemann Andersen, M. (2011) "Særskrivning og sammenskrivning i dansk". I. Schoonderbeek Hansen & P. Widell (red.): *13. Møde om Udforskningen af Dansk Sprog*. Århus (under udgivelse).

Hoaas, K. Andvik (2008) *Særskrivningsmønstre. En kvantitativ og kvalitativ studie av VG1-elevers sær- og samskrivning av sammensatte ord i norsk*. Masteroppgave ved Institutt for lingvistiske og nordiske studier. Universitetet i Oslo.

Jervelund, A. Ågerup (2007) *Sådan staver vi – om ortografi og stavefejl*. Dansklærerforeningens Forlag og Dansk Sprognævn.

Jervelund, A. Ågerup & J. Schack (2008) "På strejftog indenfor Retskrivningsordbogens territorie". A. Svavarsdóttir, G. Kvaran, G. Ingólfsson & J.H. Jónsson (red.): *Nordiske Studier i Leksikografi 9*. Reykjavík: Nordisk Forening for Leksikografi. 259-267.

Jervelund, A. Ågerup & J. Schack (2010) "En undersøgelse af elevernes stavefærdighed i FSA 2008, retskrivning". *Dansk Noter* 3, september 2010. 50-54.

Johansson, S. & Graedler A-L. (2002) *Rocka, hipt og snacksy. Om engelsk i norsk språk og samfunn*. Kristiansand: HøyskoleForlaget.

Johnson, K. (2008). Quantitative methods in linguistics. Blackwell Publishing Ltd. Malmgren, S-G. & R. Vatvedt Fjeld (2006) "Om felaktiga särskrivningar i svenskan och norskan". H. Lorentzen & L. Trap-Jensen (red.): *Nordiske Studier i Leksikografi 8*. København: Nordisk Forening for Leksikografi.

Mobärg , Mats (1997) "Om särskrivning, engelska och gestalttext". *Språkvård* 1/1997. 20-26.

Retskrivningsordbog (1955) København: Nordisk Forlag.

Retskrivningsordbogen (2001) København: Aliena – Aschehoug.

Schack, Jørgen (2011) "Adverbium + præposition – ét eller to ord?" *Nyt fra Sprognævnet* 2011/3. 1-7.

Steller, P. & K. Sørensen (1993) *Engelsk Grammatik*. Munksgaards Forlag.

Tanaka-Ishii, K. and Terada, H. (2011)"Word familiarity and frequency". *Studia Linguistica, 65:1*. 96-116. Wiley Online Library.

Tryk, H. E. (1968) "Subjective scaling of word frequency". *The American Journal of Psychology, 81:2*.170-177. JSTOR.

Vatvedt, R. Fjeld (2004) "Særskrivning av sammensatte ord i norsk og svensk. Foreløpig rapport fra et pilotprosjekt". *Ord om ord 10*. Årsskrift for leksikografi. Oslo. 14-22.

Vollan, M. (2007) ""Holdnings skapende handlings planer". Særskrivning i studenttekster". *Språknytt* 2007/4. 23-28.

Vollan, M. (2009a) "Sammensatte ord i nye norskverk". *Språkrådets skrifter 1*. Oslo: Språkrådet. 64-88.

Vollan, M. (2009b) "... i et faglig utenfra og innenfra perspektiv...". Om normering av sammensatte ord og uttrykk i norsk. H. Omdal & R. Røsstad (red.): *Språknormering – i tide og utide*? Oslo: Novus Forlag. 283-299.

Walmsness, R. (1999) *Særskrivning av sammensatte ord. En studie av feiltyper hos grunnkurselever i videregående skole*. Hovedfagsoppgave i nordisk språk og litteratur. Oslo: Universitetet i Oslo.

Walmsness, R. (2002) ”Særskrivning av sammensatte ord”. *Språknytt* 3-4. 26-29.
Zola Christensen, R. & L. Christensen (2005) *Dansk Grammatik*. Syddansk Universitetsforlag.