

## Særskrivningsfejl forudsiger yderligere sproglige problemer i avistekst

Philip Diderichsen, Sune Just Christensen og Jørgen Schack  
Dansk Sprognævn

### 1. Introduktion: ”gode og sikre sprogbrugeres skriftlige sprogbrug”

Dansk Sprognævns forskningsprojekt *Ortografisk rangering af korpustekster* udspringer af en årelang og stadig uafsluttet diskussion om hvordan man indkredser og karakteriserer den variant af dansk skriftsprog der i sprognævnsbekendtgørelsens § 1 omtales som ”gode og sikre sprogbrugeres skriftlige sprogbrug” (se Diderichsen og Schack 2015).

I vores projekt undersøger vi i første omgang om vi – med den fornødne præcision – kan finde afvigelser fra retskrivningsnormen automatisk, og dernæst om antallet af normafvigelser i en given tekst kan bruges som indikator for tekstens overordnede sproglige kvalitet.

I projektets første fase har vi undersøgt om forekomsten af entydige fejl (dvs. former som er fejl i en hvilken som helst dansk kontekst, fx ”lærene” for *lærerne* og ”hierakisk” for *hierarkisk*) øger sandsynligheden for at finde yderligere fejl i en tekst – og viser sammenhæng med tekstens kvalitet mere generelt.

Et pilotforsøg hvor Sprognævnets medarbejdere bedømte tekster med og uden sådanne fejlfund, tyder på at forekomsten af entydige fejl i en tekst øger sandsynligheden for både flere ortografiske fejl i teksten og flere øvrige sproglige problemer såsom manglende kohærens m.m., altså lavere tekstkvalitet mere generelt, uafhængigt af de ortografiske fejl (Diderichsen og Schack 2015).

I anden fase af projektet arbejder vi med sprogteknologiske metoder til at identificere betingede fejl, dvs. former som er fejl i den

konkrete kontekst, men som kan være korrekte i andre kontekster. I et andet pilotforsøg så vi således på r-fejl som *fungere* i ”Vi har en ledelse, der fungerer som et kollektiv” og *kombinerer* i ”911 Turbo har både kraften og komforten til at kombinerer supersportsvognndrømmen med rollen som hverdagsbil”. Også disse fejltyper ser ud til at øge sandsynligheden for yderligere fejl i en tekst (Diderichsen, Just Christensen og Schack 2015; under udgivelse).

I denne artikel fremlægger vi et tredje pilotforsøg hvor særskrivningsfejl er i fokus. Som det vil fremgå, er der også her en sammenhæng mellem fejl der kan fremfindes semiautomatisk, og yderligere sproglige problemer, set i forhold til sammenlignelige tekster.

## **2. Dataindsamling**

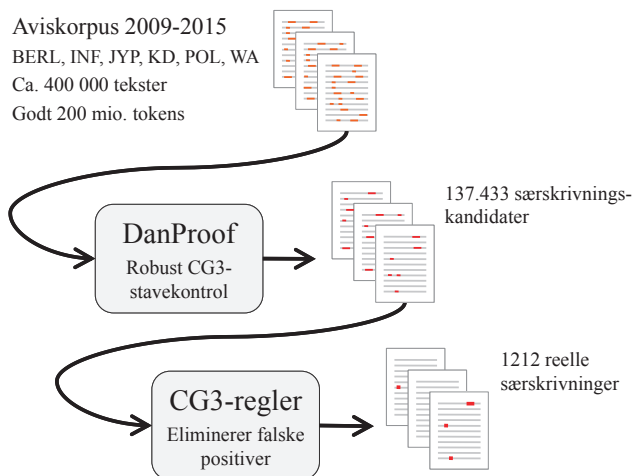
Langt det største arbejde i forbindelse med denne undersøgelse har været at finde forekomster af særskrivninger. Det er ingen sag at finde to sammenlignelige tekster – kunsten er overhovedet at finde nogen med særskrivningsfejl.

Særskrivningsfejl finder vi i vores Infomediakorpus, som er opmærket og efterfølgende stave- og grammatikkontrolleret. Det anvendte korpus består af et bredt udvalg af de artikler der er trykt i Berlingske Tidende, Information, Jyllands-Posten, Kristeligt Dagblad, Politiken og Weekendavisen fra januar 2009 til august 2015. Dette svarer til godt 200 millioner løbende ord fordelt på ca. 400.000 tekster. Teksterne er vha. stave- og grammatikkontrollen DanProof opmærket med blandt andet lemmaangivelser, ordklasse- og bøjningskategorier og markeringer af stave- og grammatikfejl (Bick 2001; 2015). Alt sammen ved hjælp af Constraint Grammar (CG), en grammatikformalisme der eliminerer sproglige flertydigheder ved ud fra konteksten at udvælge korrekte læsninger af flertydige ord og ordforbindelser. Se Bick (2015) for en mere detaljeret beskrivelse af hvordan DanProof virker samt referencer til litteratur om Constraint Grammar.

DanProof markerer hver af sine identificerede fejltyper med et fejlspecifikt mærke, og ved særskrivningsfejl skelner DanProof mellem forskellige typer ud fra om et ord bør sættes sammen med det efterfølgende eller det forudgående ord, og om sammenskrivningen

kræver bindestreg. DanProof er især god til at identificere særskrivninger der kræver bindestreg, og denne type særskrivning er derfor hyppigt repræsenteret i vores materiale. Dette er væsentligt at holde sig for øje da det særskrivningsmateriale vi ender med at finde frem til, ikke kan forventes at være repræsentativt i forhold til de særskrivningstyper der kan findes i avistekst. Der mangler m.a.o. højst sandsynligt en del ”ganske almindelige” særskrivninger (dvs. uden bindestreger) i vores materiale. Det er imidlertid ikke et metodemæssigt problem for netop denne undersøgelse da vi ser på rigtighedsprocenten for hver enkelt af de særskrivninger vi finder, og ikke for særskrivninger generelt.

Vores dataindsamlingsprocedure er kort fortalt at gennemgå alle de ordforbindelser som DanProof markerer som særskrivninger (særskrivningskandidater) og derefter at opstille nye CG-regler som kan frasortere de særskrivningskandidater som af en eller anden grund er fejlagtigt identificeret (falske positiver). Proceduren er illustreret i Figur 1 og uddybes i det følgende.



Figur 1. Forløb fra korpus til identificerede faktiske særskrivninger.

For at kunne identificere en særskrivningskandidat som enten sand eller falsk positiv må vi have klare retningslinjer for hvad vi opfatter

som særskrivninger. Det er ikke i alle tilfælde i overensstemmelse med hvad DanProof markerer som særskrivninger. Fx markerer DanProof generelt engelske flerordsudtryk der betegner stillinger (såsom *senior general manager*, *business manager* og *control manager*) som særskrivninger, men der er vi mere kritiske. Så snart (dele af) disse ordforbindelser bøjes på dansk, som fx *business manageren*, eller hvis utvetydigt danske ord gøres til en del af ordforbindelser som *business manager vejleder*, kategoriserer vi dem som særskrivninger. Men kan ordforbindelserne siges at være direkte citeret fra engelsk som faste helheder, betragter vi dem ikke som særskrivninger, men behandler dem som falske positiver. DanProof har også en tendens til at markere flerleddede *proprier* som *Mary Fonden* og *Anti Doping Danmark* som særskrivninger. Men eftersom de almindelige retskrivningsregler ikke uden videre gælder for *proprier* (RO §61.1), markerer vi ikke disse som ægte særskrivningsfejl, medmindre der da tilføjes led til ordforbindelsen som ikke er en del af *propriet* (som fx *Carlsberg Fondens ledelsen*). Andre forbindelser med *proprier* som fx *Harley Davidson motorcykel* og *Duonic automatgearkasse* hvor forbindelsens sidsteled er så tungt at det muliggør en udtale med to hovedtryk, har vi også angivet som falske positiver. En sidste type hvor vi konsekvent må afvise DanProofs særskrivningsanalyse, er ordforbindelser som syvetagers højhuse og 6-trins gearkasse, som heller ikke entydigt kan siges at være særskrevne, da dette afhænger af udtalen.

Det kan til en vis grad diskuteres hvorvidt de ovenstående ordforbindelser er eksempler på særskrivning eller ej. Men ved andre ordforbindelser som vi også har markeret som falske positiver, kan det ikke diskuteres da markeringen af ordforbindelsen som en særskrivning beror på en fejlanalyse fra DanProofs side. Fx foreslår DanProof ”Birthe Rønn Hornbech-fredag” ud fra sekvensen ”.. efter at Birthe Rønn Hornbech fredag i Politiken proklamerede ..”. Dette kan håndteres med en regel der også eliminerer andre falske positiver med ugedagenes navne såsom ”.. hun skrev i sin leder lørdag ..” m.fl. Eksempler på de reelle særskrivninger vi finder, kan ses nedenfor, efter Tabel 1.

Vi har pga. falske positiver som de ovennævnte måttet frasortere over 99 % af DanProofs særskrivningskandidater. Hermed er vi gået fra hele 137.433 særskrivningskandidater til bare 1212 faktiske særskrivninger fordelt på 944 forskellige sammensatte ord i 1012

avisartikler.

Det næste skridt var at finde ud af hvor ofte de enkelte sammensætninger egentlig staves korrekt. Det er vi interesserede i for at kunne finde ud af om der er nogen teksttræk der hænger særlig stærkt sammen med en tendens til særskrivning, fx tekstens emne (er der forskel på særskrivningsprocenten i tekster om erhverv, kultur, politik, sport?), udgivelsesår (stiger eller falder den sproglige korrekthed over årene?) eller kilde (er der forskel på de enkelte avisers skribenter og/eller korrekturkvalitet?). Det kan være relevant for hvilken type tekster vi evt. skal have ekstra fokus på i fremtiden.

For at kunne finde sær- og sammenskrivningsprocenterne i forhold til disse forskellige baggrundsvARIABLE, fandt vi frem til de korrekt sammenskrevne former af hver af de fejlagtigt særskrevne sammensætninger vi havde fundet.

Det er vigtigt at notere sig at dette er en anden måde at opgøre procenter for sær- og sammenskrivninger på end man normalt ser i litteraturen. Typisk tælles alle sammensætninger i et begrænset materiale, og det opgøres så hvor mange af dem der er særskrevne (se fx Heidemann Andersen i dette bind). Da det ikke er muligt i vores materiale at optælle samtlige sammensætninger i hånden<sup>1</sup>, går vi i stedet ud fra de særskrivninger vi identificerer, idet vi kompletterer materialet med de tilsvarende korrekte former. Det vil nødvendigvis give en vis særskrivningsbias i materialet. Det betyder at man skal være meget varsom med at sammenligne vores procenttal med procenter i andre undersøgelser. Til gengæld mener vi ikke at der er noget væsentligt til hinder for at betragte procentangivelserne i det følgende som retvisende med hensyn til relative tendenser i forhold til emne, år og kilde.

Og hvad finder vi så? Det ser vi på i det følgende.

### **3. God og sikker brug af sammensætninger**

I Tabel 1 vises 10 tilfældige sammensætninger af de 944 forskellige der findes i vores data, og deres overordnede korrekthedsprocenter.

---

<sup>1</sup> De få procent sammensætninger der findes i løbende avistekst (en hurtig optælling på en tilfældig avisside giver 4 %) bliver til adskillige millioner forekomster i vores korpus med 200 mio. løbende ord.

| Sammensætning               | Fejl | Korrekte | % korrekte |
|-----------------------------|------|----------|------------|
| BNP-vækst                   | 1    | 187      | 99,5       |
| C20-selskaber               | 1    | 50       | 98,0       |
| energiagentur               | 1    | 28       | 96,6       |
| public service-stationernes | 1    | 5        | 83,3       |
| public service-opgave       | 5    | 11       | 68,8       |
| EM-slutrundekamp            | 1    | 1        | 50,0       |
| hånd-øje-koordination       | 1    | 1        | 50,0       |
| EG-koncernen                | 2    | 0        | 0,0        |
| pro-Mubarak-demonstranter   | 2    | 0        | 0,0        |
| V8-biturbo                  | 4    | 0        | 0,0        |

Tabel 1. Korrekthedsprocenter for 10 tilfældige sammensætninger fra vores data (vist korrekt samskrevet). Stikprøven er trukket fra sammensætninger der findes mindst to gange i vores korpus.

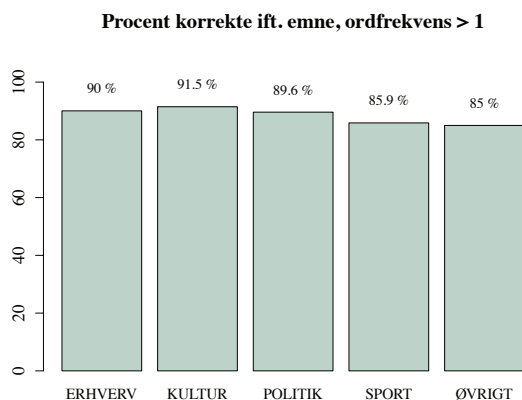
Korrekthedsprocenterne tager udgangspunkt i sammensætninger som forekommer særskrevet (jf. ovenfor). Det er derfor karakteristisk at flere af sammensætningerne slet ikke findes korrekt stavet i vores korpus på trods af at den viste stikprøve kommer fra mængden af sammensætninger der forekommer mindst to gange i korpuset.

For fuldstændighedens skyld kommer her 20 tilfældige sammensætninger der kun forekommer en enkelt gang i hele korpuset. Da vi jo tager udgangspunkt i særskrevne former, vil disse hapax legomena i sagens natur alle være særskrevne. Vi viser dem i den form de forekommer i teksterne:

*3.g pensummet, Coca-Cola drenge, favorit kylling, Fletcher grammo-fon, Gilbern racer, GTE udgaven, her-og-nu scenario, Højre backen, IQ niveau, Klima fondene, lalala-Madonna pastiche, matematik-fysik linje, Minerva bil, open-source datacenter, os-mod-verden sammenhold, Q8 crossover, sen-sæson destination, syresødme forhold, umami tricket, øre-til-øre kommunikation*

I det følgende viser vi korrekthedsprocenter i forhold til nogle af de baggrundsvARIABLE der er til rådighed i datamaterialet.

Et af de teksttræk som vi har adgang til, er teksternes emne: Avisteksterne i korpuset er på forhånd opmærket med emneord af deres skribenter. Se Figur 2.



Figur 2. Sammensætningskorrekthedsprocenter i forhold til overordnede artikelemner. Procenterne er baseret på sammensætninger der forekommer mindst to gange i vores korpus, mindst en gang særskrevet.

I Figur 2 kan man bl.a. se at tekster om sport ligger noget lavere i korrekthedsprocent end de andre. Og det ser ikke ud til at være på grund af mere komplicerede sammensætninger i sportssektionen, for de varierer ikke voldsomt mellem emnekategorierne, se Tabel 2. Det tyder på at det kan være en smule mere interessant at kigge på sportstekster end på andre tekster hvis man er interesseret i at finde fejl.

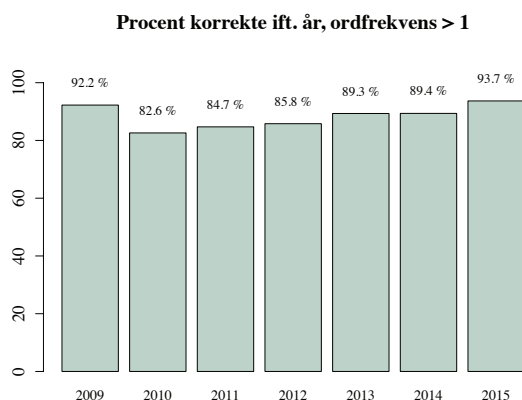
| Erhverv (n = 6924) | Kultur (n = 1008)                 | Politik (n = 15996)          | Sport (n = 4085)         | Øvrigt (n = 12662)     |
|--------------------|-----------------------------------|------------------------------|--------------------------|------------------------|
| verdensøkonomien   | samfundsdebatten                  | selskabsskatten              | VM-finalen               | VM-kvalifikation       |
| BNP-vækst          | fitnesscenter                     | espressomaskinen             | EM-finalen               | kønskampen             |
| CO2-udledning      | garagerockband                    | CO2-udledningen              | antidoping-myndighederne | arbejdsmarkedet        |
| MBA-program        | <b>først-til-mølle princippet</b> | EU-lande                     | VM-finalen               | arbejdsmarkedet        |
| arbejdsmarkedet    | public service-kontrakt           | klippekort                   | verdensranglistens       | <b>old boys hippie</b> |
| Eurozonen          | <b>drive-in biograferne</b>       | arbejdsmarkedet              | ottendedelsfinalerne     | FN-sanktionerne        |
| premium-segmentet  | realityprogrammerne               | arbejdsmarkedet              | arbejdsmarkedet          | arbejdsmarkedet        |
| selskabsskatten    | DAB-radio                         | arbejdsmarkedet              | biturbomotor             | fitnesscenter          |
| selskabsskatten    | tv-nyhederne                      | public service-forpligtelser | verdensranglisten        | tv-station             |
| EU-lande           | P1-programmet                     | euroobligationer             | ottendedelsfinalerne     | EU-formandskabet       |

Tabel 2. Stikprøver på 10 tilfældige sammensætninger fra hvert af emnerne i Figur 2. Over hver kolonne er angivet hvor mange sammensætninger der findes i tekster med det givne emne. De understregede ord er de eneste særskrivninger i stikprøven.

Bemærk at "Øvrigt"-kategorien indeholder en masse ord der ser ud til at passe helt fint ind i de fire andre kategorier. Når de alligevel er havnet i "Øvrigt"-kategorien, skyldes det at de tekster de forekommer i, ikke har fået tildelt emneord af skribenten eller andre. Desuden er det påfaldende at et ord som *arbejdsmarkedet* forekommer i adskillige af kolonnerne. Det afspejler at det er den absolut hyppigste sammensætning: 10969 af de i alt 40675 sammensætningsforekomster er *arbejdsmarkedet*. Kun én af de 10969 gange forekommer ordet særskrevet.

### *Særskrivningsfejl forudsiger yderligere sproglige problemer i avistekst*

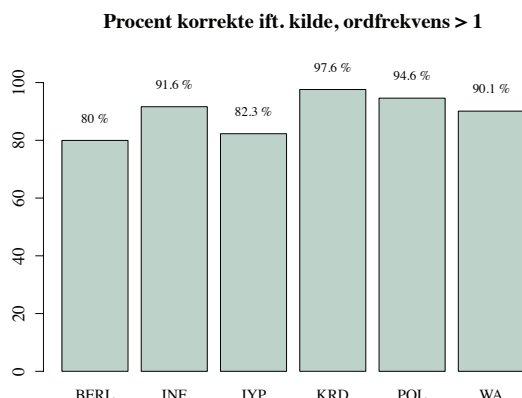
En anden måde at anskue vores data på er pr. år. I Figur 3 kan man se at korrekthedsprocenten blandt sammensætninger (der forekommer særskrevet) generelt har været stigende i de senere år.



Figur 3. Sammensætningskorrekthedsprocenter i forhold til udgivelsesår. Procenterne er baseret på sammensætninger der forekommer mindst to gange i vores korpus, mindst en gang særskrevet.

Det vil nok overraske en og anden læserbrevsskribent, men det svarer til hvad vi har set tidligere i projektet. Vi har således også kunnet konstatere et fald i antallet af entydige fejl i løbet af de senere år.

Ser vi på de forskellige kilder i vores datamateriale, viser der sig en vis variation i sammensætningskorrektheden fra avis til avis. Er man interesseret i at finde særskrivninger, vil der være lidt mere at hente i Berlingske og Jyllandsposten end i Kristeligt Dagblad.



Figur 4. Sammensætningskorrekthedsprocenter i forhold til kilde. Procenterne er baseret på sammensætninger der forekommer mindst to gange i vores korpus, mindst en gang særskrevet.

De ovenstående analyser giver et godt overblik over vores sammensætningsdata. Variationen i forhold til de forskellige tekstvariable er ikke dramatisk, men man kan alligevel ane visse mønstre i hvor den bedste og sikreste sprogbrug, i hvert fald hvad angår sammensætninger, forekommer. En mere detaljeret statistisk analyse af hvilke baggrundsvariable der er forbundet med en forøget sandsynlighed for særskrivning, vil føre for vidt her (men indgår i planerne for det videre arbejde med projektet).

I stedet vil vi nu fremlægge endnu en pilotundersøgelse af sammenhængen mellem ortografiske fejl fundet i konkrete enkelttekster og yderligere sproglige problemer i disse tekster.

#### **4. Særskrivninger forudsiger yderligere fejl**

Hvis vi skal kunne tro på at det kan lade sig gøre at udvikle en metode der kan finde frem til teksters relative sproglige kvalitet ved hjælp af

semiautomatisk identifikation af fejl, skal vi helst kunne vise at tekster der indeholder identificerede fejl, også generelt har en lavere kvalitet end tilsvarende tekster uden identificerede fejl. I de pilotforsøg vi allerede kort har omtalt ovenfor, har dette vist sig at holde stik.

Det første af disse pilotforsøg tog udgangspunkt i tre tekster der indeholdt entydige fejl og tre matchende tekster (se Diderichsen og Schack 2015). I to ud af tre tekstpar fandt vi den forventede sammenhæng, altså flere yderligere fejl i fejlteksterne end i matchteksterne.<sup>2</sup>

Det andet pilotforsøg tog udgangspunkt i fire andre tekster med r-fejl og entydige fejl. Her fandt vi den forventede sammenhæng i 3 ud af 4 tekstpar<sup>3</sup> (Diderichsen, Just Christensen og Schack 2015; under udgivelse).

Generelt tyder vores hidtidige resultater altså på at ortografiske fejl forudsiger yderligere sproglige problemer. Nu er spørgsmålet så om det samme mønster går igen i vores særskrivningsdata.

For at undersøge dette tilrettelagde vi en ny pilotundersøgelse der skulle vise om tekster med særskrivningsfejl identificeret af os ville indeholde flere yderligere sproglige problemer end matchende tekster uden særskrivninger identificeret af os. Undersøgelsen forløb som følger.

Vi udvalgte fem tekster blandt dem der indeholdt flest særskrivninger, se Tabel 3. (Bemærk de mange biltekster, der hører under det overordnede emne sport).

---

<sup>2</sup> I det tredje tekstpar var matchteksten uden entydige fejl tilfældigvis fyldt med særskrivningsfejl! Hvis vi på det tidspunkt havde identificeret de særskrivninger vi har nu, var den valgte matchtekst ikke blevet valgt.

<sup>3</sup> Også her var der en ekstraordinær fejlkilde på spil: Det tekstpar der ikke udviste sammenhængen, viste sig at indeholde en fejlidentificeret r-fejl, nemlig ”ophøre”, der stod i et citat fra 1735:

”I 1735 skrev han [Otto Thott, deputeret i Kommercekollegiet], at det var et problem, at nogle kræmmere ”aldrig snarere (har) vunden 20 á 30.000 rigsdaler, og vel mindre, førend de ophøre at handle, og da søger characterer for sig”.”

| Nr. | Kilde           | År   | Emne         | Særskrivninger                                                                                                            |
|-----|-----------------|------|--------------|---------------------------------------------------------------------------------------------------------------------------|
| 1.  | Berlingske      | 2011 | Biler        | lithium-ion batterier, lithium-ion batterier, Plug-in hybrid                                                              |
| 2.  | Information     | 2011 | Mediepolitik | fundraiser direktøren, public service radiostationer, public service radiostationer                                       |
| 3.  | Berlingske      | 2014 | Biler        | V6 benzin, Puttgarden-Rødby færgen, CO2 udledning                                                                         |
| 4.  | Berlingske      | 2014 | Mediepolitik | public service forpligtelser, public service kanaler, public service opgave, public service opgave, public service opgave |
| 5.  | Jyllands-Posten | 2015 | Biler        | science fiction bil, GT-R modellen, science fiction bil                                                                   |

Tabel 3. Fem udvalgte fejltekster og de semiautomatisk identificerede særskrivninger i dem.

De fem fejltekster sammenlignede vi med fem matchtekster uden identificerede særskrivninger. Disse tekster fandt vi ved at matche på kilde, år, emner, forfatterkøn og tekstlængde. Vi sammenlignede altså fem par af tekster der mindede om hinanden bortset fra at vi havde fundet særskrivninger i den ene, ikke i den anden.

Vi (artiklens forfattere) gennemgik herefter hver for sig hver af de 10 tekster for sproglige problemer guidet af Ole Tøgebys stilistikmanual *Stiltræk* (Tøgeby 1996). Som illustrativt eksempel kommer her en gennemgang af hvilke problemer vi fandt i tekstpar 4.

Det fremgår af Tabel 3 hvilke særskrivninger vi fandt i fejltekst 4. I Tabel 4 og Tabel 5 kan man se hvilke yderligere fejl og problemer vi fandt i de to tekster. Vi fandt forekomster af forkert stavede navne, misbrug af bindestreger, samskrivningsfejl, (yderligere) særskrivningsfejl, apostroffejl, manglende ord, kongruensfejl, kommafejl, problemer med opbygning af konstruktioner (herunder ordstilingsfejl) og sammenblanding af faste udtryk. Og som man kan se, er der flest problemer i fejlteksten.

*Særskrivningsfejl forudsiger yderligere sproglige problemer i avistekst*

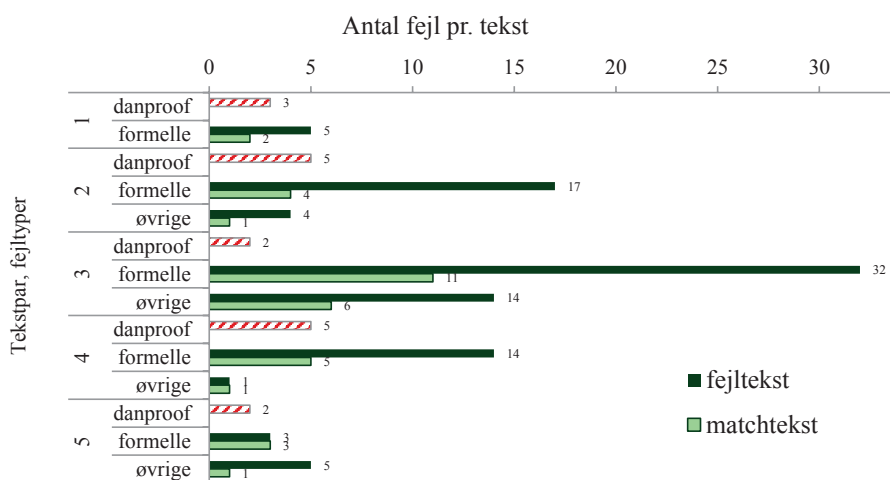
| Fejl/problem                                                                                                                                                                                | Korrekt form/kommentar                                                       |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| maggieterning                                                                                                                                                                               | Det hedder Maggi ..                                                          |
| drama-serier                                                                                                                                                                                | Misbrug af bindestreg                                                        |
| public serviceaftale                                                                                                                                                                        | Samskrivning                                                                 |
| 70'erne                                                                                                                                                                                     | 70'erne                                                                      |
| 80'erne                                                                                                                                                                                     | 80'erne                                                                      |
| DRK                                                                                                                                                                                         | Samskrivning                                                                 |
| fire-årig                                                                                                                                                                                   | Misbrug af bindestreg                                                        |
| .. frem for nogle af public service forpligtelser ..                                                                                                                                        | .. af de public ..                                                           |
| nogle af mest populære programtyper                                                                                                                                                         | nogle af de mest populære ..                                                 |
| medieforsker Københavns Universitet                                                                                                                                                         | medieforsker ved ..                                                          |
| en helt kanal                                                                                                                                                                               | en hel kanal                                                                 |
| ,                                                                                                                                                                                           | Ekstra komma                                                                 |
| Men et stigende antal seere benytter sig af muligheden for via nettet eller deres TV-distributør som for eksempel YouSee at se udsendelserne, når de har tid og lyst – såkaldt on demand-TV | Ophobning af led. Bl.a. ”som for eksempel YouSee” bidrager til kluntetheden. |

Tabel 4. (Yderligere) sproglige fejl og problemer i fejlttekst 4.

| Fejl/problem                                           | Korrekt form/kommentar                                          |
|--------------------------------------------------------|-----------------------------------------------------------------|
| public service aftale                                  | Særskrivning                                                    |
| at DR med sin store organisation og mange penge        | .. og sine mange ..                                             |
| at DR ganske kraftigt sine opruster online aktiviteter | .. opruster sine ..                                             |
| public servicestation                                  | Samskrivning                                                    |
| en ny tredje platform                                  | en ny, tredje platform                                          |
| forrest i fødekæden                                    | Mulig kontamination af ”forrest i køen” og ”øverst i fødekæden” |

Tabel 5. (Yderligere) sproglige fejl og problemer i matchtekst 4.

Dette mønster går igen i alle tekstparrene, som man kan se i Figur 5. De særskrivninger vi har fundet automatisk, er vist i en separat bjælke (med etiketten ”danproof”). Derudover vises optællinger af formelle fejl (ortografiske og grammatiske fejl, manglende ord) og øvrige problemer (kohærensproblemer o.l.). I hvert tekstpar er det tydeligt at fejltjeksterne indeholder flere problemer end matchteksterne – selv når man ser bort fra de fejl vi havde identificeret på forhånd.



Figur 5. Antal fejl/sproglige problemer i de undersøgte fejl- og matchtekster. De skraverede bjælker viser antallet af særskrivningsfejl identificeret af os i de enkelte fejltjekster ud fra DanProofs særskrivningskandidater.

Mønsteret for særskrivninger ligner således de resultater vi tidligere har fået: Automatisk identificerede fejl forudsiger yderligere sproglige problemer i vores korpus. Det virker derfor mere og mere sandsynligt at korpus tekster kan kvalitetsrangeres efter antallet af ortografiske fejl.

## **5. Konklusioner**

For at kunne rangere korpustekster ortografisk er vi nødt til at kunne regne med at forekomsten af ortografiske fejl forudsiger yderligere sproglige problemer. Resultaterne fra særskrivningsundersøgelsen er en yderligere indikation af at dette faktisk er tilfældet generelt.

I vores korpus af ca. 400.000 avistekster fra danske dagblade i perioden 2009 til 2015 fandt vi 1212 særskrivninger fordelt på 944 forskellige sammensatte ord i 1012 artikler. En sammenligning af fejltekster og matchtekster viste hvad vi også tidligere har set: Tekster med fundne betingede fejl (her: særskrivninger) forudsiger yderligere fejl og generelt lavere tekstkvalitet end tekster uden fundne betingede fejl. Vi har også set tegn på at relativt lavere tekstkvalitet tenderer til at forekomme hyppigere i visse emner (sport), år (tidligere år) og kilder (Berlingske, Jyllands-Posten). Forskellene er ikke store, men kan alligevel have en vis relevans for hvilke tekster der skal tillægges hvilken vægt i arbejdet med retskrivningsnormeringen.

Det går selvfølgelig ikke an at generalisere og erklære fx alle sportsjournalister for ”mindre gode og sikre sprogbrugere”; for det første har redaktører og korrekturlæsere et medansvar for avisernes tekster, og for det andet ligger rigtighedsprocenten for ikke-unikke særskrivninger (som i øvrigt sandsynligvis ligger i den svære ende pga. den nævnte overvægt af bindestregssammensætninger) i sportsteksterne kun 5,6 procentpoint under den bedste kategori, kulturteksterne. Men vi har dokumenteret at hvis man er interesseret i at finde særskrivningsfejl, vil man kunne finde en smule flere i sportstekster end i tekster om kultur, erhverv og politik, og at sandsynligheden for tekster af en lidt lavere kvalitet generelt således også vil være en smule højere blandt sportstekster.

Vi har nu ad flere omgange set lignende resultater: At ortografiske fejl forudsiger yderligere sproglige problemer, selv i et korpus af tekster af generelt høj kvalitet (avistekster). Projektet fortsætter derfor nu med systematisk indsamling af flere velkendte betingede fejltyper således at vi i sidste ende kan sammenligne et delkorpus med alle de mest gængse ortografiske fejl med et tilsvarende, fejlfattigt delkorpus med henblik på ortografisk variation og normering. Her skal man fortsat holde sig for øje at disse gængse ortografiske fejl vil blive brugt

som kvalitetsindikatorer på samme måde som særskrivningerne i den ovenfor beskrevne undersøgelse. Der er altså ikke tale om at gængse fejl vil kunne godkendes. Det er derimod tanken at disse fejl kan hjælpe os med at frasortere tekster præget af mindre sikker sprogbrug således at nye stavevarianter der kan være betinget af usikkerhed, kan forkastes i normeringsarbejdet.

## **Litteratur**

- Bick, Eckhard (2001) En Constraint Grammar Parser for Dansk. 8. *Møde om Udforskningen af Dansk Sprog*, 12.-13. oktober 2000:40-50
- Bick, Eckhard (2015) DanProof: Pedagogical Spell and Grammar Checking for Danish. *Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, Sep 7-9 2015*:55-62.
- Diderichsen, Philip & Jørgen Schack (2015) Jagten på den gode og sikre sprogbruger. *Nyt fra Sprognævnet* 2015, 3:1-8.
- Diderichsen, Philip, Sune Just Christensen & Jørgen Schack (2015) Ranking corpus texts by spelling error rate. *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)*. Francesco Mambrini et al. (udg.). Warszawa, Polen: Institute of Computer Science, Polish Academy of Sciences. 31-38.
- Diderichsen, Philip, Sune Just Christensen & Jørgen Schack (under udg.). Ortografisk rangering: At ordne tekster efter antallet af stavfejl. *Studier i Nordisk* 2012-2015.
- Heidemann Andersen, Margrethe (2017) Sprognormer på nettet. 16. *Møde om Udforskningen af Dansk Sprog*, 13.-14. oktober 2016:21-37.
- RO = Anita Ågerup Jervelund et al. (udg.) (2012) *Retskrivningsordbogen*. 4. udg. København: Alinea.
- Togeb, Ole (1996) *Stiltræk*. Århus: Det Humanistiske Fakultet, Aarhus Universitet.