Yonatan Goldshtein, Inger Schoonderbeek Hansen
og Tina Thode Hougaard (udg.):
17. Møde om Udforskningen af Dansk Sprog, Århus 2018

Nye danske kerneord − kerneordskorpus 2.0

Philip Diderichsen og Jørgen Schack
Dansk Sprognævn

1. Introduktion

Det umiddelbare formål med projektet *Nye danske kerneord*[1] er at opbygge et tekstkorpus, *kerneordskorpus 2.0*, der gør det muligt at fremstille en nutidig frekvensordliste som er direkte sammenlignelig med en ældre frekvensordliste, nemlig Hanne Ruus' liste over danske kerneord, som omfatter perioden 1970-1974 (Ruus 1995). Sammenligningen af de to ordlister skal i første omgang danne grundlaget for en beskrivelse af udviklingen i arten og antallet af moderne kerneimportord ud fra to fikspunkter, nemlig de to femårsperioder 1970-1974 og 2012-2016 (jf. Schack 2019).

Når man laver sprogforandringsundersøgelser, fx en ordforråds-undersøgelse som den netop omtalte, må man sørge for at de to synkrone snit som holdes op imod hinanden, faktisk *er* sammenlignelige. I vores tilfælde er der som minimum to krav som skal være opfyldt. For det første er det "ikke nok at have to korpusser hvis tekster stammer fra to forskellige tidspunkter. Korpusserne skal også være sammensat af netop de samme tekstarter i netop det samme blandingsforhold" (Duncker 2013:161). For det andet skal de lemmaer der optræder på den nye ordliste, være udvalgt efter nøjagtig samme kriterier som dem der optræder på den ældre liste. Der skal m.a.o. benyttes helt ensartede lemmatiseringskriterier (jf. afsnit 5).

[1] Projektet indgår i det overordnede projekt *Yes, det er coolt. Om påvirkning af dansk fra andre sprog* med det formål at undersøge antallet af moderne importord blandt hyppige ord. Se Schack (2019) samt Heidemann Andersen (dette bind) og Diderichsen (dette bind).

2. Hvad er et kerneord?

De to ordinventarer vi skal sammenligne, består hver især af såkaldte kerneord. Et kerneord (eller mere teknisk: et kernelemma) er ifølge Ruus (1995:38) et lemma der forekommer "meget hyppigt" i et korpus der er baseret på den af de undersøgte tekstarter (jf. afsnit 3) der indeholder det mindst påfaldende (dvs. mest almene) ordforråd. Denne definition forudsætter at man på en pålidelig måde kan beregne om et lemma er mere hyppigt end et andet lemma. Her udnyttes den antagelse at et lemmas forekomsttal er poissonfordelt over korpusser, hvilket meget groft sagt betyder at et lemma der forekommer fx 4 gange i et givet korpus, også vil forekomme i omegnen af 4 gange i andre, tilsvarende korpusser. Et lemma er således (i forhold til poissonfordelingen) signifikant hyppigere end et andet lemma, hvis man kan trække to standardafvigelser (dvs. 2 gange kvadratroden af forekomsttallet) fra det første lemmas forekomsttal uden at nå ned på det andet lemmas forekomsttal (Maegaard & Ruus 1986:10 f.). Forekomstgraden "meget hyppigt" definerer Ruus på følgende måde: "Ved **meget hyppigt** vil jeg forstå **meget mere hyppigt**, idet jeg betragter et lemma som **meget hyppigt**, hvis det er mere hyppigt end et lemma, som er mere hyppigt end et lemma, som med sikkerhed optræder i en tekstart" (ibid.). Denne lidt kringlede men meget præcise definition er lettere at forstå hvis man ser på et konkret eksempel (tabel 1).

3	4	11	**20**
ikke signifikant forskelligt fra 0	signifikant hyppigere end 0	signifikant hyppigere end 4	signifikant hyppigere end 11
panel	*tape* (sb.)	*turist*	*tjekke* (vb.)

Tabel 1. Illustration af forekomstgraden "meget hyppigt"

I kerneordskorpus 2.0 forekommer ordet *tape* (sb.) 4 gange. 4 forekomster er (i modsætning til 3 forekomster og derunder) signifikant forskelligt fra 0 forekomster, og *tape* er derfor et ord der statistisk set med sikkerhed forekommer i den undersøgte tekstart. Ordet *turist* forekommer 11 gange og er dermed hyppigere end *tape*. Og nu er

vi fremme ved sagens kerne: Ordet *tjekke* (vb.), der forekommer 20 gange, er mere hyppigt end *turist*, som er mere hyppigt end *tape*, og dermed er *tjekke* et kerneord. Et lemma er altså et kernelemma hvis det forekommer mindst 20 gange i den undersøgte tekstart.

3. Korpusopbygning efter DANwORD-principperne

Ruus' kerneordsundersøgelse ligger i direkte forlængelse af Maegaard og Ruus' DANwORD-projekt, der opgjorde ordfrekvenser i 5 korpusser baseret på 5 forskellige tekstarter: ugeblade, aviser, børnebøger, romaner og fagblade (Maegaard og Ruus 1986:5 ff.). Hvert af disse 5 korpusser består af 1000 tekstprøver a 250 ord, dvs. i alt 250.000 ord. Disse er fordelt på 50.000 ord pr. år i femårsperioden 1970-1974. Tekstprøverne er udtaget fra de mest læste tekster i de enkelte tekstarter.

En sammenligning af ordforrådet i de 5 korpusser viste at de meget hyppige ordformer fra ugebladene gennemgående har en større spredning end tilsvarende ordformer fra de øvrige tekstarter (Ruus 1995:39). Frekvenslisten baseret på ugebladstekstprøver er altså den af de 5 frekvenslister der har det mindst påfaldende ordforråd, og det er derfor ugebladsordlisten der er lagt til grund for den efterfølgende udtagelse af kerneord (jf. tabel 2).

År	**Ugeblade**	Aviser	Børnebøger	Romaner	Fagblade	Alle
2012	**50.000**	50.000	50.000	50.000	50.000	250.000
2013	**50.000**	50.000	50.000	50.000	50.000	250.000
2014	**50.000**	50.000	50.000	50.000	50.000	250.000
2015	**50.000**	50.000	50.000	50.000	50.000	250.000
2016	**50.000**	50.000	50.000	50.000	50.000	250.000
Total	**250.000**	250.000	250.000	250.000	250.000	1.250.000

Tabel 2. Kerneordskorpus 2.0: ugeblade 2012-2016

Kerneordskorpus 2.0 er som nævnt sammensat på nøjagtig samme måde som kerneordskorpus 1.0 (dvs. Maegaard og Ruus' ugebladskorpus) og består derfor af 1000 tekstprøver a 250 ord, fordelt på 50.000 ord pr. år i en femårsperiode, nemlig femårsperioden

2012-2016. Korpusset er sammensat sådan at de mest læste tekster bidrager med flest tekstprøver. Læsertallene stammer fra Index Danmark/Gallup, hvis læsertalsopgørelser bygger på ca. 24.000 årlige interviews (https://danskemedier.dk/maalinger/laesertal).

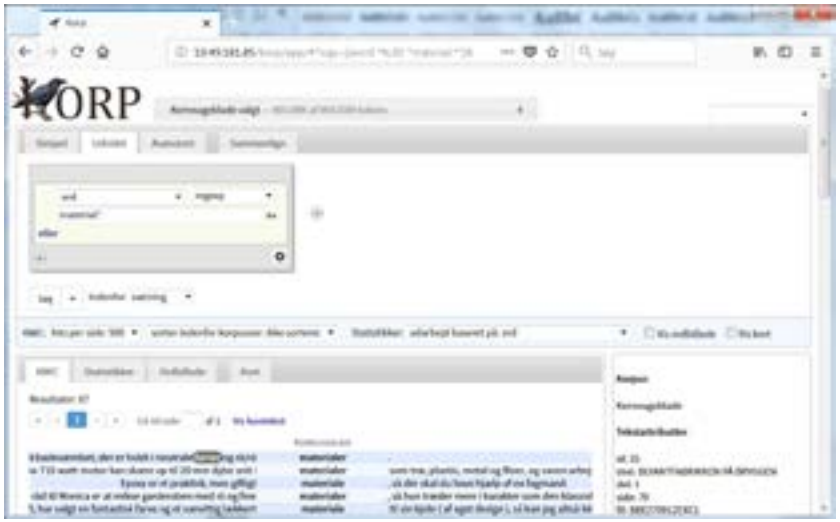
4. Opbygning af kerneordskorpus 2.0 og udtræk af nye kerneord

I praksis formede vores fremfinding af nye kerneord sig som den følgende procedure med to overordnede trin: 1) opbygning af nyt kerneordskorpus og 2) udtræk af nye kerneord.

Opbygningen af korpusset indebar først og fremmest et overraskende stort arbejde med at skaffe de nødvendige ugeblads-tekster og bringe dem på digital form i en acceptabel kvalitet. Det indebar diverse tekstoprensningsopgaver fra semiautomatisk korrektur af besværlige ligaturfejl (typografiske sammentrækninger af fx ”fi” bliver nogle gange til ”i” på vejen fra layoutet side til ren digital tekst) over korrektur af diverse andre tekstfejl (fx ÅRETS *naRRøv*) og registrering af stavefejl (fx *legendraiske*) til komplet indtastning af ikkekonventionelle tekster såsom taleboblerne fra Anders And & Co.

Efter dette store arbejde blev teksterne opmærket med den automatiske tagger/parser DanGram (Bick 2001) og renset for systematiske uoverensstemmelser i lemmatiseringen (se nedenfor). I den opmærkede tekstmasse på i alt 903.075 ord blev der herefter ved en automatisk, randomiseret procedure udpeget 1000 tekstprøver a 250 ord og altså således et kerneordsmateriale i form af en undermængde af det samlede korpus på i alt 250.000 ord.

Det samlede tekstmateriale på 903.075 ord inklusive de udpegede tekstprøver på i alt 250.000 ord blev herefter indlæst i korpusdatabasen CWB (Corpus Workbench) og gjort tilgængeligt som et korpus i webgrænsefladen Korp, se Figur 1 (Borin et al. (2012); se også Språkbankens offentlige implementering på https://spraakbanken.gu.se/korp).



Figur 1. Kerneordskorpusset (ugeblade) i grænsefladen Korp

Herefter blev kerneordene udtrukket af det færdige korpus efter en separat procedure. Et datasæt blev udtrukket med hvert ordtoken og dets DanGram-tilskrevne lemmaform og ordklasse, og ud fra dette datasæt blev der produceret et aggregeret datasæt med antallet af de enkelte lemmaer mv. Det aggregerede datasæt måtte til slut håndkorrigeres for ikke-systematiske uoverensstemmelser i lemmatiseringen (se afsnit 5).

De forskellige rettelser af DanGrams automatiske lemmatisering der var nødvendige, beskrives i det følgende.

5. Erfaringer fra håndkorrigeret autolemmatisering

Ruus har i sin kerneordsundersøgelse benyttet en rent formel lemmadefinition der (med enkelte modifikationer) svarer til den der kan uddrages af Retskrivningsordbogen, 1. udgave, 1986. Ifølge denne lemmadefinition, der er uændret i de følgende udgaver af ordbogen, er opdelingen i opslagsord ”principielt uafhængig af ordenes betydning. Det bevirker, at ord med forskellig betydning er slået sammen i ét opslagsord, hvis de i øvrigt har samme stavemåde, udtale, ordklasse og bøjning, og hvis de indgår i sammensætninger på samme måde” (Retskrivningsordbogen 1986:16). Ved lemmatiseringen af

kerneordskorpus 2.0 har vi af hensyn til sammenligneligheden måttet benytte samme rent formelle princip. En konsekvent efterlevelse af dette princip kan medføre at nogle meget hyppigt forekommende lemmaer ikke opnår status som kernelemmaer fordi det ikke er muligt på formelt grundlag at udskille de 20 belæg som kerneordskriteriet kræver. Det drejer sig især om homografe lemmaer med samme ordklasse, fx 1. *sol* sb. (et himmellegeme) overfor 2. *sol* sb. (en opløsning) og 1. *æg* sb. (den skarpe kant på en kniv e.l.) overfor 2. *æg* sb. (fx hønseæg). Oplagte kerneordskandidater som 1. *sol* og 2. *æg* risikerer at udgå af listen fordi de lemmaadskillende bøjningsformer, fx pluralisformen *sole* og singularisformen *ægget*, ikke forekommer i tilstrækkeligt omfang i det undersøgte korpus. Det nytter ikke at der er tilstrækkeligt med polylemmatiske ordformer der ikke kan adskille lemmaerne, fx *solen* og *æggene* (jf. Ruus 1995:115 ff.).

Det er desuden et problem ved sådanne polylemmatiske former at selv mindre fejlfordelinger af ordformer på de forskellige mulige lemmaer (i den automatiske lemmatisering) kan føre til at ord der egentlig er kerneord, ikke bliver identificeret som sådanne.

Der er også andre lemmatiseringsproblemer der ikke kan løses automatisk, og som kan være vanskelige for selv de skarpeste filologer. Det giver vi et eksempel på i næste afsnit.

5.1 Ikke-systematiske uoverensstemmelser mellem DanGram og RO 2012

I dette afsnit giver vi et enkelt eksempel på et særlig drilsk lemmatiseringsproblem som ikke kan løses automatisk.

Lemmaet *hotel* med ordklassen substantiv har 18 forekomster. Forekomsterne er for størstedelens vedkommende upåfaldende, se eksemplerne nedenfor.[2] I *Hotel d'Angleterre* er *Hotel* lemmatiseret korrekt som substantiv, jf. Ruus (1995:113): "Da hele arbejdet med at finde kerneord bygger på enkelte ordformer, er proprier fra konkordanserne bestående af flere ordformer ikke registreret som sådanne."

[2] Dog er der en enkelt ting der springer i øjnene, nemlig en enkelt brug af *Hotel* med stort begyndelsesbogstav (korrekt iflg. RO § 12.7, jf. fx *Region Syddanmark*) som en del af navnet *Hotel d'Angleterre* (i øvrigt i modsætning til "Tjæreborgs hotel Stella Polaris", hvor *hotel* ikke er en del af navnet, men fungerer som en appellativisk kategoribetegnelse.

optaget på Costa del Sol og Tjæreborgs	**hotel**	Stella Polaris.
Holdt op imod de	**hoteller**	, jeg allerede havde stiftet bekendtskab med ..
Kort efter ankomsten til	**Hotel**	D'Angleterre torsdag eftermiddag begyndte der ..
De ville gerne betale for en uge eller to på et	**hotel**	, og allerede samme dag overførte de et beløb ..
.., men byens	**hoteller**	har ofte gode tilbud i weekenderne ..

Tabel 3. Lemma: hotel, ordklasse: substantiv. Uddrag af konkordans på 18 forekomster

De 18 forekomster af former af *hotel* er umiddelbart for lidt til at dette lemma kan tælle med som kerneord.

Vi kan imidlertid se i vores datamateriale at DanGram-taggeren ikke konsekvent holder sig til den ene ordklasse *hotel* kan have iflg. Retskrivningsordbogen, og *Hotel d'Angleterre* ser faktisk ud til at være en afvigelse fra DanGrams normale måde at håndtere den slags eksempler på.

En konkordans af lemmaet *hotel* hvor ordklassen ifølge DanGram ikke er et substantiv, men derimod (en del af) et proprium, giver 10 forekomster af *Hotel*/*HOTEL* med stort begyndelsesbogstav med et hotelnavn enten før eller efter – eller begge dele:

Placering af hotelnavn	Forekomst	Antal
Hotelnavn før *Hotel*	Paradise Hotel	3
	COMWELL HOTEL	1
	Crawford Hotel	1
	Beverly Hilton Hotel	1
Hotelnavn efter *Hotel*	Hotel D'Angleterre	1
	Hotel Plaza	1
	Hotel Skt. Petri	1
Hotelnavn både før og efter	Grand Hotel Quisisana	1

Tabel 4. Forekomster af lemmaet hotel med ordklassen proprium iflg. DanGram

Med 10 ekstra forekomster af lemmaet *hotel* (af DanGram kategoriseret som proprier) ud over de 18 DanGram havde identificeret som substantiver, burde lemmaets status som kerneord nu være genoprettet med i alt 28 forekomster, der alle ifølge RO er substantiver.

Der er dog endnu en lille spidsfindighed der skal tages højde for, nemlig ordenes eventuelle status som udenlandske ord. De 6 forekomster af *Hotel* med et navn forrest virker alle udenlandske. Selve navnene indbyder til en engelsk udtale (*Paradise*, *Crawford*, *Beverly Hilton*, *COMWELL*), og sammenstillingen med *Hotel* udgør i alle tilfælde tydeligt ét proprium med flere ordformer, hvorved det naturligste bliver også at bruge en engelsk udtale på *Hotel*. Noget tilsvarende gør sig gældende for *Grand Hotel Quisisana*, her dog måske snarere med en fransk eller italiensk udtale (hotellet ligger på Capri). Hvis disse former skal klassificeres som udenlandske ord, kan de ikke bidrage til at hæve antallet af *hotel*, sb.-forekomster op på de kerneordskritiske 20.

I så fald er der kun navnene på de danske hoteller tilbage: *Hotel D'Angleterre*, *Hotel Plaza* og *Hotel Skt. Petri*. Her vil den mest sandsynlige udtale af *hotel* i alle tre tilfælde være dansk, og dermed ender vi trods alt med 3 ekstra forekomster af *hotel* ud over de 18 identificeret som substantiver af DanGram og således i alt 21 – lige akkurat nok til at lemmaet bliver et kerneord.

Man kan diskutere om henvisning til udtale som ovenfor er rimeligt indenfor den strengt formelle lemmatisering der er lagt op til i Ruus (1995). Hvis udtaleovervejelser som de ovenstående ikke må tages i betragtning, er der på den anden side ingen måde at identificere, endsige forkaste de udenlandske forekomster af *hotel* på – formen *hotel* er monolemmatisk i Retskrivningsordbogen, og samtlige forekomster vil i så fald automatisk skulle tilskrives ordklassen substantiv. I tilfældet *hotel* bliver resultatet således det samme uanset hvad: Ordet bliver et kerneord.

Der er mange ting der således må håndteres manuelt. Dog er der en del lemmatiseringsproblemer der er systematiske og kan løses automatisk.

5.2 Systematiske uoverensstemmelser mellem DanGram og RO 2012
Ved en gennemgang af en tidlig version af kerneordslisten identificerede vi et antal systematiske uoverensstemmelser mellem DanGrams grammatiske opmærkning og Retskrivningsordbogen. Der er dels tale om uoverensstemmelser i ordklasser der synes at være udtryk for en anden, men i og for sig rimelig grammatisk synsvinkel, dels om deciderede fejlanalyser. Vi giver eksempler på disse to typer uoverensstemmelser i dette afsnit.

Et udpluk af systematiske ordklasseuoverensstemmelser er vist i tabel 5.

Eksempel	Lemma	DanGram-ordklasse	Korrekt ordklasse iflg. RO 2012
Senere *fik hun tegnet endnu flere ..*	sen	adv.	adj.
I ***begge*** *ender af hver gjord ..*	begge	determinativ	pron.
.. 11 år yngre ***end*** *sin kæreste ..*	end	præp.	konj.
Bering House ***of*** *Flowers*	of	præp.	(udenl. ord)

Tabel 5. Eksempler på systematisk afvigende ordklassetilskrivninger i DanGram

Lemmaet *sen* klassificeres af DanGram som et adverbium når det står adverbielt i sætningen – en helt igennem rimelig analyse, men en ordklassetilskrivning der er i uoverensstemmelse med Retskrivningsordbogen, hvor *sen* kun har ordklassen adjektiv. (Når *sen* står som adjektiv, klassificerer DanGram det som et sådant, fx *en* ***sen*** *aften*, *de* ***seneste*** *år* osv.). Pronomenet *begge* klassificeres konsekvent som determinativ af DanGram. Determinativ er en syntaktisk kategori, ikke en ordklasse, så her har den grammatiske funktion igen fået forrang i DanGrams klassificering (ligesom den adverbielle funktion af *sen* får forrang for ordklassen adjektiv). Analysen er igen i og for sig rimelig, men ifølge Retskrivningsordbogen har *begge* ordklassen pronomen, og kun den. Konjunktionen *end* klassificeres af DanGram

som en præposition hvis det efterfølges af en nominalhelhed, men som en underordningskonjunktion hvis det efterfølges af en sætning. Igen en helt rimelig analyse på et funktionelt niveau, men ifølge Retskrivningsordbogen kan *end* ikke have ordklassen præposition, kun konjunktion eller adverbium. Et ganske særligt tilfælde udgøres af kategorien "udenlandske ord", dvs. fremmed ordstof der ikke har status som fremmedord, og som derfor ikke er med i de ordbøger der beskriver det danske ordforråd, inkl. Retskrivningsordbogen. I praksis drejer det sig kun om de engelske ord *of* og *the*, der indgår i et stort antal engelsksprogede navne og titler. DanGram tildeler disse ord den ordklasse de har i engelsk, og i kerneordssammenhæng skal de klassificeres som "udenlandske ord", uden ordklasseangivelse (jf. Ruus 1995:114).

Ud af de ca. 1200 lemmaer på vores første tentative kerneordsliste har vi indtil videre fundet 75 sådanne systematiske uoverensstemmelser, altså ca. 6 %. De fordeler sig på korrekte ordklasser ifølge Retskrivningsordbogen som vist tabel 6.

DanGram-ordklasse rettes til ...	**Antal lemmaer**
adjektiv	29
pronomen	15
substantiv	10
verbum	8
konjunktion	4
udenlandsk ord	3
præposition	3
adverbium	2
talord	1

Tabel 6. Optælling af hvor mange lemmaer der skal tilskrives end anden ordklasse end den DanGram tilskriver

Det er især adjektiver der er blevet fejlklassificeret af DanGram, efterfulgt af pronomener, substantiver, verber og et antal småordklasser. Pronomenerne udgøres i de fleste tilfælde af lemmaer der klassificeres som determinativ: *begge*, *denne*, *du*, *enhver*, *hvilken*, *ingen*, *jeg*, *nogen*, *selv*, *sin*. Det kan undre at de personlige pronomener *du* og *jeg* optræder her. Det skyldes at *din* og *min*, der i Retskrivningsordbogen og den grammatiske tradition i øvrigt klassificeres som possessive pronomener, af DanGram klassificeres som genitivformer af de personlige pronomener *du* og *jeg*. Hvorfor *selv* klassificeres som determinativ, er der ingen oplagt forklaring på givet eksempler som fx *sådan en kan man da* ***selv*** *lave*, *det må du* ***selv*** *se*, *Jeg synes jo ikke* ***selv****, jeg er lækker*.

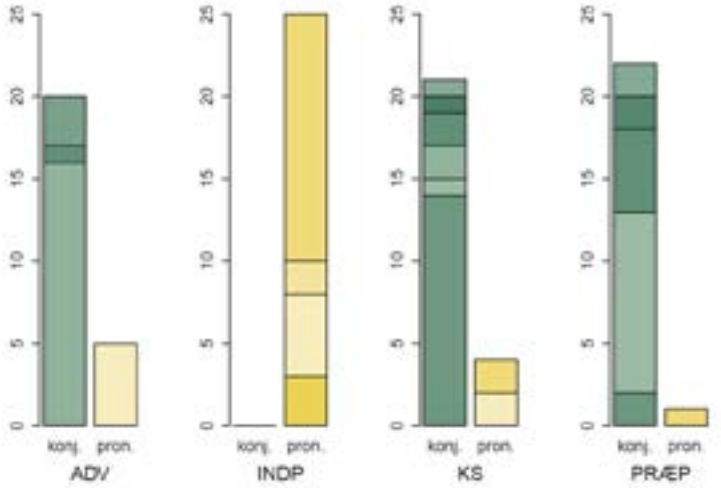
Et eksempel på en decideret systematisk fejlanalyse fra DanGrams side kan ses i klassifikationen af ordformen *siden*. I tabel 7 vises DanGrams klassificering af 812 forekomster af *siden*[3] underopdelt efter de korrekte ordklasser. Den næsten helt systematiske fejlanalyse kan ses i DanGrams klassificering af *siden* i sætningsinitial position som bestemt form af substantivet *side*. Her skal hele 95,5 % af forekomsterne i stedet klassificeres som adverbier. Her drejer det sig om eksempler som ***Siden*** *har hun kæmpet for erstatning og retfærdighed*, ***Siden*** *havde ingen set dem* og lignende. Den resterende ene forekomst er en præposition: ***Siden*** *da har jeg været allergisk.* I dette lille hjørne af den danske grammatik er der plads til forbedring af DanGrams regler, som vil kunne indfange den slags forekomster i en klart bedre analyse uanset grammatisk synsvinkel. Også de 23,1 % sætningsinitiale præpositioner klassificeret som adverbier af DanGram vil kunne korrigeres systematisk: Der er i alle 6 tilfælde tale om forbindelsen *siden da* eller *siden dengang*, som i ***Siden da*** *er der sket meget på markedet*. Også andre af de særligt hyppige fejlklassifikationer af *siden* vil kunne korrigeres systematisk.

[3] Fundet i det komplette korpus, altså også i de dele af teksterne der ikke er udvalgt til tekstprøver.

DanGram-klassificering af *siden* fordelt på korrekte ordklasser		Ikke-initial placering	Initial placering	Total
adv. (DanGram)		**377**	**26**	**403**
	adv.	99,2 %	76,9 %	97,8 %
	konj.	0,3 %	0,0 %	0,2 %
	præp.	0,3 %	23,1 %	1,7 %
	sb.	0,3 %	0,0 %	0,2 %
konj. (DanGram)		**84**	**10**	**94**
	adv.	9,5 %	0,0 %	8,5 %
	konj.	85,7 %	100,0 %	87,2 %
	præp.	4,8 %	0,0 %	4,3 %
præp. (DanGram)		**100**	**15**	**115**
	adv.	13,0 %	0,0 %	11,3 %
	konj.	1,0 %	0,0 %	0,9 %
	præp.	86,0 %	100,0 %	87,8 %
sb. (DanGram)		**178**	**22**	**200**
	adv.	19,1 %	95,5 %	27,5 %
	konj.	0,6 %	0,0 %	0,5 %
	præp.	5,6 %	4,5 %	5,5 %
	sb.	74,7 %	0,0 %	66,5 %

Tabel 7. DanGrams klassificering af ordformen *siden*. Tallene er delt op efter om *siden* forekommer sætningsinitialt (som i Siden har der været stille) eller ej (som i Det var længe siden ...). Tal i halvfed er rå forekomsttal, resten er procenter af disse

Et andet eksempel er DanGrams klassificering af *som*, se Figur 2.



Figur 2. DanGrams klassificering af ordformen *som*. DanGrams ordklasser ADV, INDP, KS og PRÆP er fordelt på Den Danske Ordbogs to ordklasser konjunktion og pronomen

Figur 2 viser 25 tilfældigt udvalgte forekomster af *som* af DanGram klassificeret som ”ADV”, 25 klassificeret som ”INDP” (”independent pronoun”, altså pronomen), 25 som ”KS” (”[k]onjunction, subordinating” (sic.), altså konjunktion) og 25 som ”PRÆP”. Indenfor hver af DanGrams ordklasser er forekomsterne underopdelt efter om de svarer til Den Danske Ordbogs (DDO, se ordnet.dk) to hovedbetydninger af *som*, hhv. den konjunktionale og den pronominale betydning. De mørke nuancer i venstresøjlerne svarer til diverse underbetydninger af konjunktionsbetydningen i DDO, de lyse nuancer i højresøjlerne svarer til *som*s forskellige grammatiske

funktioner i den relativsætning det indleder (hhv. subjekt, objekt, styrelse og knudesubjekt). Det væsentlige er det overordnede billede: DanGram klarer den overordnede klassifikation af *som* som pronomen godt (alle DanGrams INDP'er svarer til den ældre traditions analyse af *som* i funktionen som relativt pronomen) mens resten af DanGrams kategorier svarer nogenlunde til DDO's konjunktion. Imidlertid havde DanGram helt kunnet undgå fejlklassificeringer ved at følge Retskrivningsordbogen, der følger den nyere danske grammatiske tradition, hvor *som* klassificeres som konjunktion og kun som konjunktion. At ordbogen måske skulle overveje at tilføje ordklassen præposition, er en helt anden sag (jf. klart præpositionelle anvendelser som *Ham er du da mindst lige så god som*).

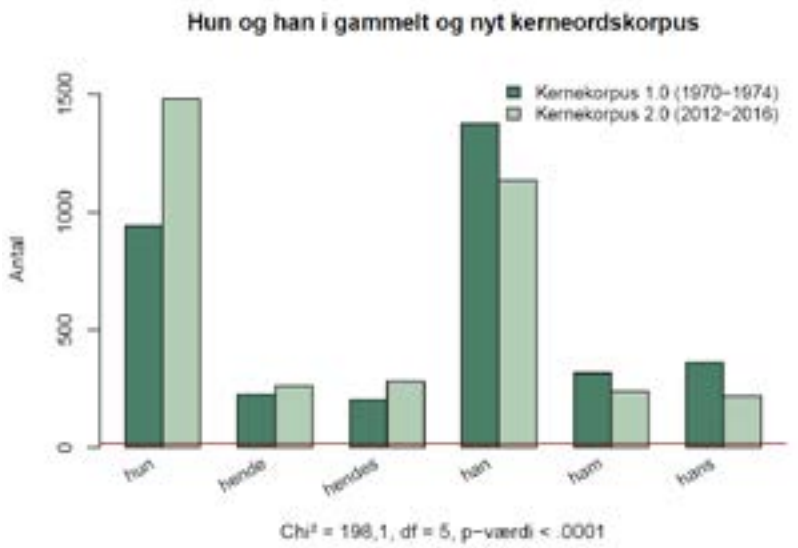
Den minutiøse kuratering af kerneordene fører til mange overvejelser af denne art, hvor det kan være svært at træffe en helt sikker afgørelse. Trods al mulig metodestringens vil man derfor næppe helt kunne undgå subjektivt farvede kategoriseringer i fremfindingen af kerneordene.

6. Kerneordskorpus 1.0 vs. 2.0 – et par diakrone observationer

Kerneordskorpus 2.0 vil kunne bruges til meget pålidelige undersøgelser af diakrone ændringer i det centrale ordforråd. Vi foretager her et par første sammenlignende nedslag i det gamle og det nye kerneordskorpus.

6.1 Hun vs. han

I Figur 3 ses et diagram over udviklingen fra det gamle kerneordskorpus til det nye af pronomenerne *hun* og *han*. De enkelte former af de to lemmaer er vist så man kan se den konsekvente udvikling der er sket.



Figur 3. Hyppigheder af lemmaerne *hun* og *han* inkl. bøjningsformer i det gamle og det nye kerneordskorpus. Den vandrette linje repræsenterer kerneordsgrænsen på 20 forekomster

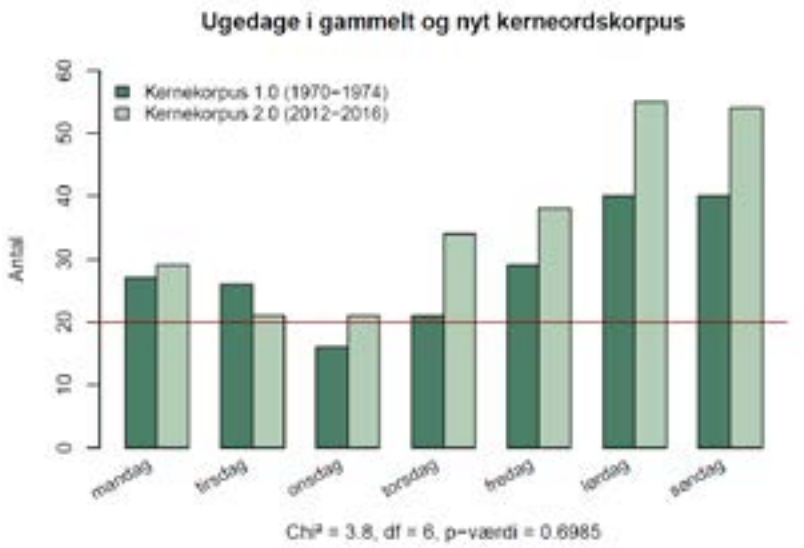
De mørke søjler i Figur 3 repræsenterer hyppighederne i det gamle kerneordskorpus, de lyse søjler hyppighederne i det nye. I venstre side af diagrammet kan man således se at hyppigheden af *hun* er steget med over 50 %. Formerne *hende* og *hendes* er ligeledes steget en smule. I højre side kan man derimod se at hyppigheden af *han*, *ham* og *hans* er faldet noget. Det samlede antal af lemmaerne *hun* og *han* er kun steget med knap 6 % i forhold til det gamle kerneordskorpus (fra 3427 til 3623 forekomster), og den markante udvikling i brugsmønsteret er da også statistisk højsignifikant ($Chi^2 = 198$, $df = 5$, $p < 0.0001$). Ordformen *hun* i det nye korpus har sågar overhalet *han* i det gamle. Der er her selvfølgelig ikke tale om ændringer i sammensætningen af det centrale ordforråd – *hun* og *han* ligger begge langt over

kerneordsgrænsen på 20 forekomster (markeret med vandret linje i Figur 3). Men vi har her at gøre med en diakron ændringsobservation (vel at mærke i "virkelig tid"[4]) der formentlig er så pålidelig som den overhovedet kan blive. Det synes indiskutabelt at kvinder omtales hyppigere i ugebladene nu end tidligere. Og hvad bunder denne ændring i brugsmønsteret så i? Man kan umiddelbart forestille sig to overordnede muligheder: 1) Der er kommet flere artikler om og med kvinder; 2) kvinderne er kommet mere på banen i de enkelte artikler – eller en kombination af disse. Uanset hvordan det således ser ud rent tekststrukturelt, er det nok så relevant at spørge til indholdssiden: Er ændringen udtryk for at der omtales flere kvinder i færd med at gøre traditionelle 'kvindeting', eller er der snarere tale om at kvinder i højere grad er trængt ind på mændenes traditionelle domæner? Vi kan ikke komme det nærmere på det foreliggende grundlag, navnlig fordi teksterne fra det gamle kerneordskorpus ikke opbevares samlet længere. Men man vil kunne undersøge det nærmere ved at gennemlæse en stikprøve af relevante ugeblade fra de to perioder 1970-1974 og 2012-2016.

6.2 Ugedagene

I forbindelse med udgivelsen af Danske Kerneord (Ruus 1995) var der en del der morede sig over at alle ugedagene var kerneord – undtagen *onsdag*. Vi har sammenlignet hyppighederne af ugedagene i det gamle og det nye korpus. Resultatet kan ses i figur 4.

[4] Indenfor sociolingvistiske sprogforandringsstudier opererer man med studier i hhv. forestillet og virkelig tid efter William Labovs begreber *apparent time* og *real time* (Gregersen & Kristiansen 2015:12 f. og 33 f.). Studier i forestillet tid indebærer at man undersøger sproget hos informanter i forskellige aldersklasser ud fra den antagelse at forskellene er udtryk for igangværende sprogforandring. Studier i virkelig tid er undersøgelser gentaget med år(tier)s mellemrum. Studier i virkelig tid deles op i panelstudier, hvor sproget fra den samme flok informanter underkastes sådanne gentagne studier, og trendstudier, hvor sprog der så vidt muligt kommer fra tilsvarende kilder, undersøges med passende mellemrum – som her.



Figur 4. Hyppigheder af ugedagene i det gamle og det nye kerneordskorpus. Den vandrette linje repræsenterer kerneordsgrænsen på 20 forekomster

Det viser sig at alle ugedagene i denne omgang kvalificerer sig som kerneord – også *onsdag*.

Hvad der imidlertid er væsentligere, er at brugsmønsteret i dette lukkede leksikalske paradigme ikke har ændret sig i statistisk signifikant grad fra dengang til nu ($Chi^2 = 3{,}8$, df = 6, p = 0,6985). Det overordnede forekomsttal for ugedagene er ganske vist steget med over 25 % i forhold til det oprindelige kerneordskorpus (fra 199 til 252 forekomster), og det forklarer hvordan *onsdag* nu lige nøjagtig formår at opnå kerneordsstatus. Men brugsmønsteret ugedagene imellem er altså det samme, med flere referencer til weekenddagene og færre til dagene i starten og især midten af ugen. Det virker ikke rimeligt at konkludere at der er sket væsentlige ændringer i det centrale ordforråd. Således har *onsdag* i virkeligheden samme status som dengang, uanset om det har været kerneord eller ej.

7. Konklusion og perspektiver

Kerneordskorpus 2.0 er et nyt ugebladskorpus med materiale fra 2012-2016 som er maksimalt sammenligneligt med det oprindelige kerneordskorpus fra 1970-1974. Vi har her beskrevet sammensætningen og opbygningen af korpusset, herunder de lemmatiseringsproblemer der skal håndteres i forbindelse med automatisk grammatisk opmærkning. En konklusion er at alt ikke går af sig selv bare fordi man benytter sig af automatiserede metoder. Der er i allerhøjeste grad stadig brug for lingvistisk ekspertviden i arbejdet med et sådant korpus. Vi vil også fremhæve oprensningen og kurateringen af det digitale tekstmateriale vi har arbejdet med – den alene har kostet projektet anseelige ressourcer. Derudover er der som beskrevet systematiske og ikke-systematiske uoverensstemmelser mellem den automatiske opmærkning og den grammatiske norm vi arbejder indenfor. Empiri tager tid – også med hjælpemotor.

Vi har i et par nedslag demonstreret hvordan det nye kerneordskorpus kan bruges til meget pålidelige diakrone undersøgelser af ordforrådet i samspil med hyppighedslisterne fra det gamle korpus: *Hun* har taget førertrøjen fra *han* i ugebladene, og *onsdag* har samme status som det havde dengang – uanset om det har været kerneord eller ej.

Der er oplagte fremtidsperspektiver i det nye korpus på kortere og længere sigt. Først og fremmest skal korpussets oprindelige formål selvfølgelig opfyldes og evt. nye moderne importord i kerneordforrådet identificeres. Derefter vil det være oplagt at få lavet en detaljeret analyse af tilvæksten af nye kerneord i det hele taget – og bortfaldet af gamle kerneord. En interessant korpuslingvistisk undersøgelse vil være at finde ud af hvor vigtigt det egentlig er at kuratere et korpus til mindste detalje som det er gjort her. Hvor stor forskel gør det egentlig? Ville man evt. kunne 'nøjes' med et større, men til gengæld mindre velkontrolleret korpus?

I modsætning til det oprindelige kerneordskorpus har vi med Kerneordskorpus 2.0 ikke bare en optælling af lemmaerne, men et levende korpus indlæst i en brugervenlig søgegrænseflade (Korp), hvor lemmaerne til enhver tid kan tjekkes og nye korpusundersøgelser foretages. Med de grundige retteprocedurer korpusset har været igennem, vil det kunne tjene som en form for guldstandard for

grammatisk opmærket tekst og dermed kunne bruges til nøjagtige grammatiske analyser af sproglige fænomener, så længe de har en vis hyppighed. Korpusset er jo nemlig ikke voldsomt stort, en egenskab der i sig selv kan gøre det til en mere overskuelig opgave løbende at rette den grammatiske opmærkning og på den måde måske endda gøre korpusset endnu bedre over tid.

Tak

Forfatterne vil gerne takke Hanne Ruus, som har ydet en uvurderlig hjælp i forbindelse med opbygningen af kerneordskorpus 2.0. Vi takker også de studentermedhjælpere der var med til at indsamle, rense og forberede teksterne i korpusset: Anders Schack, Winnie Collin, Nanna Hansen, Amalie Arleth Viinholt og Lasse Halberg Haarbye.

Litteratur

Andersen, M. Heidemann (2019) Yes, det er coolt. Moderne importord i danske aviser. Y.U. Goldshtein, I. Schoonderbeek Hansen & T. Thode Hougaard (red.) *17. Møde om Udforskningen af Dansk Sprog*. Århus: Aarhus Universitet:45-57.

Bick, Eckhard (2001) En Constraint Grammar Parser for Dansk. Peter Widell & Mette Kunøe (red.) *8. Møde om Udforskningen af Dansk Sprog, 12.-13. oktober 2000*. Århus: Århus Universitet:40-50.

Borin, Lars, Markus Forsberg & Johan Roxendal (2012) Korp – the corpus infrastructure of Språkbanken. *LREC*:474-78.

Diderichsen, Philip (2019) Moderne importord over tid i høj opløsning. Y.U. Goldshtein, I. Schoonderbeek Hansen & T. Thode Hougaard (red.) *17. Møde om Udforskningen af Dansk Sprog*. Århus: Aarhus Universitet:231-255.

Duncker, Dorthe (2013) Kerneord og DANwORD som grundlagsmetode for korpusprofilering. Dorthe Duncker mfl. (red.) *Betydning og forståelse. Festskrift til Hanne Ruus*. København: Selskab for Nordisk Filologi.

Gregersen, F. & T. Kristiansen (2015) Indledning: Sprogforandring i virkelig tid. Gregersen, F. & Kristiansen, T. (red.) *Hvad ved vi nu – om danske talesprog?* København:Sprogforandringscentret.

Maegaard, Bente og Hanne Ruus (1986) *Hyppige ord i danske aviser, ugeblade og fagblade*. Bind 2. København: Gyldendal.

Retskrivningsordbogen (1986) Dansk Sprognævn (udg.) København: Gyldendal:16.
RO = *Retskrivningsordbogen*, https://ds.dk/ro.
Ruus, Hanne (1995) *Danske Kerneord. Centrale dele af den danske leksikalske norm* 1-2. København: Museum Tusculanums Forlag.
Schack, Jørgen (2019) Kerneimportord. *Yes, det er coolt. Moderne importord i dansk* [under udgivelse].

Sidder, går og står og … : et interaktionelt blik på kongruenskonstruktion i dansk

Yonatan Goldshtein
Aarhus Universitet

1. Indledning

I denne artikel undersøges en grammatisk konstruktion i dansk på baggrund af dens brug i naturligt forekommende samtale. Undersøgelsen baserer sig teoretisk på den interaktionelle lingvistik, der søger at undersøge hvordan sprog faktisk bruges af talere i naturlig tale (Couper-Kuhlen &Selting 2018). Der er sket meget i beskrivelsen af det danske samtalesprogs grammatik siden oprettelsen af forskningsgruppen DanTIN og deres online platform Samtalegrammatik.dk for fem år siden (Steensig et al. 2013). Selvom grammatiske ”kerneområder” som syntaks (fx Mikkelsen 2010) og morfologi (fx Hamann et al. 2012) har fået noget opmærksomhed, har størstedelen af bidragene til samtalegrammatikken dog være inden for områder, der tidligere har været stort set uberørte i traditionel grammatisk beskrivelse, så som interjektioner, partikler og enkeltordsytringer (fx Brøcker et al. 2012, Tholstrup 2014). Når grammatiske konstruktioner er blevet behandlet, har det været relativt faste konstruktioner som ”at have haft ringet” (Brøcker 2015) eller ”hvad hedder det” (Clausen & Pedersen 2017).

Konstruktionen, der her undersøges, hører til en konstruktionstype, der kaldes ”kongruenskonstruktion” (Hansen & Heltoft 2011:979-1009), og mere specifikt undergruppen, der af Nielsen (2011) kaldes ”situativ konstruktion”. Konstruktionen består på overfladen af to verber (V1 og V2), der er sideordnede med *og*, fx