Novus forlag · eISSN 1890-5455

Maal og minne 2-2023 257–303

# Grammatiske fingeravtrykk

## Normklynger og forfatterattribusjon

Helge Dyvik

Normklynger er grupper av tekster som viser felles valg blant valgfrie alternativer innenfor en skriftspråksnorm og danner 'klynger' innenfor rommet av valgmuligheter fordi andre valgkombinasjoner er sjeldnere. Basert på materiale på bokmål fra den norske trebanken NorGramBank undersøker denne studien forekomsten av normklynger som omfatter både morfologiske og syntaktiske fenomener, og muligheten for å identifisere forfatteren bak en tekst på grunnlag av plassering av forfatterens tekster i forhold til slike klynger. I avsnittene 2–3 undersøkes korrelasjoner mellom åtte grammatiske fenomener blant 338 forfattere, parallelt med en kasusstudie av forfatterattribusjon basert på én forfatter. Attribusjonsstudien utvides i avsnitt 4 gjennom sammenligning med ytterligere ni tilfeldig valgte forfattere.

Studien dokumenterer korrelasjoner mellom morfologiske og syntaktiske fenomener. Kasusstudien demonstrerer muligheten for at en forfatter på grunnlag av et lavt antall norm- og stilvalgrelaterte egenskaper noen ganger kan identifiseres blant flere hundre andre som forfatter av sine tekster. Studien av ni ytterligere forfattere støtter antagelsen om en sammenheng mellom denne muligheten og forfatterens plassering i forhold til normklynger, men indikerer at også andre faktorer spiller inn.

### 1 Innledning*

#### *1.1 Om å telle i tekster*

Grensen mellom den kvalitative og den kvantitative kunnskapstilegnelse krysses ikke alltid ustraffet. På området skjønnlitteratur, objektet *par ex-*

* Forfatteren takker Stig Jarle Helset, Victoria Rosén og to anonyme fagfeller for svært nyttige kommentarer, endringsforslag og rettelser til tidligere versjoner av artikkelen.

*cellence* for den kvalitative tilnærming gjennom tolkning, innlevelse og forståelse, må den som gir seg til å telle, regne med hevede øyenbryn. Særlig var det tilfellet for noen tiår siden. For eksempel ble en filolog med en artikkel som tellet plantebetegnelser i Henrik Ibsens lyrikk (Iversen 1944), gjenstand for André Bjerkes bitende ironi (Bjerke 1962):

> Man synes å se ham under arbeidet. Lorgnetten kommer på, og bd. XIV slåes opp; professoren tilegner seg poesi — dvs. han *teller*. Den forskende pekefingernegl, fortrolig med folianter, arbeider seg nedover sidene. Eureka! — der har vi «figenblad»: kryss i notatboken. Under rubrikk D: «Plantedeler». (Bjerke 1962:111f.)

Bjerke angrep med dette ikke bare den nevnte artikkelen, men også en tendens han mente å se innenfor humaniora mer allment – en utillatelig kryssing av grenser mellom vitenskapstyper:

> Det er selve den humanistiske forsknings misère vi ser blottlagt: dens tragikomiske forsøk på å imitere den klassiske naturvitenskap. (Ibid.:116)

Den kritiserte artikkelen har riktignok sine ufrivillig underholdende sider. Men de senere årtiers utvikling av omfattende språkressurser i form av tekstkorpora, trebanker og grammatiske analyseverktøy har gjort det stadig tydeligere at mye kunnskap om språklig variasjon og forfatteres språklige og stilistiske særpreg faktisk kan utvinnes statistisk. Både vår evne til å telle raskt i store tekstmengder og vår evne til automatisk å finne språklige egenskaper som det kan være verd å telle, har vokst i overveldende grad siden Bjerkes tid. Dermed er det også blitt mulig å oppdage hvordan tilstrekkelig store tall kan avsløre interessante mønstre i tekster. For språklig tekst er ikke bare uforutsigelig kreativitet; tekster har også kvantifiserbare mønstre og formaliserbare grammatiske strukturer. Denne dobbeltheten er en kjerneegenskap ved menneskelig språk. Naturlige språk er både uovertrufne redskaper for fantasi og kreativitet, for tenkning, kunstnerisk utfoldelse og skapelse av ny mening, og samtidig regelbundne inntil matematisk formaliserbarhet i form av komplekse regelsystemer for syntaks og morfologi. Langt fra å stå i en paradoksal motsetning til hverandre er disse to sidene intimt forbundet: Det er nettopp den komplekse regelmessighet i språket som muliggjør en nyansert og åpen diskurs og en kunstnerisk lek med språkets muligheter. Regelmessighet er uunnværlig fordi formidling av informasjon og mening le-

vende vesener imellom alltid forutsetter gjenkjennbarhet i signalet. Med den avanserte bruk menneskelige språk har, må de gjenkjennbare sidene av det talte og skrevne 'signal' nødvendigvis bli både abstrakte og komplekse, slik for eksempel syntaktiske mønstre er. Det er nødvendig hvis signalsystemet skal kunne bære en språkbruk som i tillegg til å være åpen og kreativ også er forståelig for andre.

Arbeidet bak den studien som skal presenteres her, og bak de språkressursene den bygger på, omfatter både en utførlig formalisering av de syntaktiske reglene i norsk skriftspråk, anvendt for datamaskinell analyse av tekster, og en kvantitativ behandling av de mønstrene vi deretter kan finne i de analyserte tekstene. Studien bringer to problemområder sammen: **normklynger** og **forfatterattribusjon**. Formålet er for det første å undersøke muligheten for å utvide studiet av normklynger i norsk fra morfologi og ortografi til syntaks, og for det annet å gjennomføre en kasusstudie, og en oppfølgende komparativ studie, for å belyse i hvilken grad informasjon om slike klynger er relevant for forsøk på å identifisere forfatteren bak en tekst.

*1.2 Normklynger*

Begrepet 'normklynge' er aktuelt i språk som er normert med valgfrihet av et visst omfang. Det beskriver en gruppe av tekster som treffer samsvarende valg blant normens alternativer, og som er mange nok til dermed å danne en klynge innenfor det rommet av muligheter som valgfriheten skaper. Begrepet ('norm cluster') innføres og utdypes med eksempler av Dyvik (2012), se også Helset (2017, 2018).

De offisielle normene for bokmål og nynorsk preges av omfattende valgfrihet. Denne valgfriheten beskrives i stor grad bare på ord- eller ordklassenivå; alternative former av ulike ord kan offisielt stort sett kombineres fritt i én og samme tekst (se Rosén 2000). (Ett av få unntak er reglene for valg av infinitivsform på *-e* eller *-a* i nynorsk). Derimot utelukkes i praksis mange slike offisielt mulige kombinasjoner i den **operative** norm. Språkets operative norm er dets uformulerte, faktiske norm som styrer språkbruken i tekstene og språkbrukernes vurderinger – den operative norm er språket selv, kunne man si, internalisert av lesende og skrivende mennesker.[1] I den operative norm er det i stor grad tale om

1. Se Dyvik (2003) om begrepet 'operativ norm', som ligger nær begrepene 'kvalifikasjonsnorm' (Sundby 1974) og 'internalisert norm' (Vannebo 1980, Vikør 2007).

valg mellom overlappende **subnormer**, der valgene på ordnivå begrenser hverandre (se Omdal & Vikør 2002:15). Dette betyr at tekster i betydelig grad danner **klynger** i det mangedimensjonale rommet av offisiell valgfrihet på ordnivå: I noen regioner i dette rommet utenom klyngene vil det være få eller ingen tekster. For eksempel vil få eller ingen tekster på bokmål kombinere *a*-endelse i verbformer som *kasta* og *henta* med *en*-endelse i substantiver som *gaten* og *boken*, selv om den offisielle normen ikke uttrykker noen slike begrensninger.

*1.3 Forfatterattribusjon*

Forfatterattribusjon går ut på å forsøke å avsløre hvem som har skrevet bestemte tekster. Slike forsøk har funnet sted siden 1800-tallet, med ulike metoder.[2] Et eksempel fra Norge er Geir Kjetsaas statistisk basert studie av forfatterskapet til *Stille flyter Don* (Kjetsaa 1984). I 1964 leverte Frederick Mosteller og David L. Wallace et epokegjørende bidrag til en gammel diskusjon om forfatterskapet til *The Federalist Papers*[3], basert på statistikk over forekomstene av et lite antall ord (Mosteller & Wallace 1964). Siden da har de dominerende metodene for forfatterattribusjon vært kvantitative og komputasjonelle, særlig basert på **stilometri**, som betegner bruk av statistikk for å oppnå en karakteristikk av en forfatters stil. Dette innebærer statistisk informasjon om språklige trekk som for eksempel setningslengde, ordlengde, ordfrekvenser, bokstavfrekvenser og ordforråd. Et eksempel er et kapittel i Blatt (2017), der han tester en formel som sammenholder frekvensene av 250 spesifikke ord i 600 engelskspråklige bøker av 50 forfattere. Bøkene ansees etter tur for å ha ukjent forfatter og testes mot samtlige andre bøker av samme forfatter og bøkene av de øvrige 49 forfatterne. Metoden gir korrekt forfatterattribusjon i over 99,4 % av tilfellene. De senere års utvikling av store tekstkorpora og kraftige databehandlingsteknikker har gjort denne tilnærmingen stadig mer effektiv og aktuell, og har også muliggjort statistikk over mer komplekse språktrekk enn de nevnte.

Både normklynger og forfatterattribusjon innebærer en kartlegging av forfatteres kvantifiserbare språktrekk, og er på den måten beslektede

2. Stamatatos (2009) gir en oversikt. Se også Blatt (2017).
3. *The Federalist Papers* er en serie på 146 politiske essay, der to forfattere, Alexander Hamilton og James Madison, begge hevdet at de hadde skrevet tolv av dem.

problemområder. Ved normklynger er språktrekkene begrenset til *forfatteres valg blant alternativer innenfor en norm*. De spørsmålene vi ønsker å belyse, er for det første i hvilken grad morfologiske normvalg i bokmål er korrelert med syntaktiske valg, og for det annet, gjennom en kasusstudie og en oppfølgende studie av ytterligere ni forfattere, i hvilken grad den variasjonen som genereres av den spesielle norske skriftspråksituasjonen, tillater unik identifikasjon av forfattere. Det vil da særlig være forfattere *utenom* én eller flere av normklyngene – 'normklyngeavvikerne' – som fremstår som lovende attribusjonskandidater, ettersom de befinner seg i et tynnere befolket område av normrommet[4] og dermed har færrest nærliggende konkurrenter blant de øvrige forfatterne.

Etter en presentasjon av åtte morfologiske og syntaktiske trekk i avsnitt 2 skal vi i avsnitt 3 ta for oss noen eksempler på klynger av disse trekkene, og underveis, som et eksempel, gjennomføre en kasusstudie av mulighetene for forfatterattribusjon basert på én utvalgt forfatter som er representert med seks tekster: *Dag Solstad*. Flere forfattere undersøkes i avsnitt 4. Materialet for undersøkelsen er hentet fra trebanken NorGramBank.

### *1.4 Trebanken NorGramBank*

En trebank er et syntaktisk analysert tekstkorpus – et korpus der hver setning er forsynt med en syntaktisk (noen ganger også en semantisk) analyse. Analyseformatene kan variere, og trebanker kan være blitt bygget opp gjennom rent manuell analyse, gjennom manuell analyse i kombinasjon med større eller mindre innslag av automatisk analyse, eller gjennom rent automatisk analyse (parsing).

Trebanker tillater søk etter og telling av syntaktiske egenskaper i tekster og gjør det dermed mulig å utvide studiet av normklynger fra ortografi og morfologi til syntaks. Siden syntaks ikke normeres offisielt, er dette også å utvide perspektivet for den normen som definerer rommet av muligheter, fra preskriptiv, offisiell norm til operativ, faktisk norm.[5] Spørsmålet ved denne perspektivutvidelsen blir da: *I hvilken grad er morfologiske normvalg korrelert med syntaktiske norm- eller stilvalg?* I hvilken grad finner

4. Begrepet om et 'normrom' utdypes i avsnitt 4.2.
5. Ettersom begrepet 'norm' oftest assosieres med den preskriptive, kan det riktignok noen ganger virke mindre treffende å omtale alternative syntaktiske uttrykksmåter som 'normvalg'; 'stilvalg' eller 'registervalg' kan da være bedre uttrykk.

vi normklynger av morfologi og syntaks? De syntaktiske valgene som diskuteres i de følgende avsnittene, som for eksempel enkel bestemthet, foranstilt possessiv og s-passiv, tilskrives ofte stilistiske egenskaper som også assosieres med en konservativ morfologi: en formell, konservativ stil. Å lete etter korrelasjoner mellom slike morfologiske og syntaktiske trekk innebærer derfor en viss test av antagelsene om felles stilistiske egenskaper ved de to typene trekk.

*NorGramBank* (se Dyvik & al. 2016) er en norsk trebank som er utviklet ved INESS- prosjektet[6] ved Universitetet i Bergen (se Rosén & al. (2012) og prosjektets webside http://clarino.uib.no/iness). Setningene i NorGramBank er analysert automatisk ved hjelp av den komputasjonelle norske grammatikken NorGram, utviklet gjennom flere år ved Universitetet i Bergen og basert på den syntaktiske teorien leksikalsk-funksjonell grammatikk (LFG).

Ved automatisk parsing får setninger ofte mange alternative analyser. Analysen registrert i trebanken er i over 99 % av tilfellene valgt ut på statistisk grunnlag (stokastisk disambiguering), basert på en statistisk modell utviklet på grunnlag av et manuelt disambiguert materiale. En test av 500 tilfeldig utvalgte setninger blant dem som har fått analyse, referert i Dyvik & al. (2016), indikerer at korrekt analyse av alle deler av setningen er funnet (før disambiguering) i ca. 85 % av setningene.

I denne studien skal vi ikke gå nærmere inn på trebankens analyser, men henviser isteden til arbeider som kan illustrere arten av grammatisk informasjon og graden av detalj i analysene, for eksempel Rosén & al. (2020).

NorGramBank inneholder nå ca. 160 millioner ord analysert tekst (hvorav ca. 150 millioner på bokmål), som omfatter aviser, sakprosa, skjønnlitteratur, stortingsforhandlinger og enkelte andre teksttyper i mindre omfang. Det skjønnlitterære materialet og mye av sakprosaen har prosjektet mottatt i OCR-skannet[7] form fra Nasjonalbiblioteket. I denne studien begrenser vi oss til bokmål, til tekster med kjent forfatter, og til forfattere som er representert med minst 4000 fullstendig analyserte setninger hver i trebanken.[8] Blant disse forfatterne begrenser vi

6. INESS står for 'INfrastructure for the Exploration of Syntax and Semantics'.
7. OCR står for 'optical character recognition' og innebærer mekanisk overføring av bilder av trykt eller håndskrevet tekst til maskinlesbar form.
8. Vi ser også bort fra et mindre antall deltrebanker i NorGramBank av tekniske

oss videre slik at vi får like mange forfattere av hvert kjønn. Dette gir 338 forfattere av de over 2700 som finnes i NorGramBank (se navnelisten i Appendiks, avsnitt 6.1). I de følgende avsnitt skal vi se på fordelingen av et utvalg grammatiske fenomener blant disse 338 forfatterne.

Ettersom tekstmengde er kriteriet for utvalget av forfattere, og dessuten skjønnlitteratur er en betydelig del av NorGramBank, består tekstgrunnlaget for studien hovedsakelig av romaner og, i mindre omfang, sakprosabøker. Stortingsforhandlinger er ikke inkludert.

Hovedspørsmålene for studien kan oppsummeres slik:

> Er det mulig å dokumentere norm- og stilklynger som bringer morfologiske valg fra offisiell norm sammen med syntaktiske valg fra operativ norm? I hvilken grad vil plasseringen av en forfatter i periferien av eller utenfor én eller flere slike klynger kunne bidra til korrekt attribusjon av forfatterens tekster?

Formålene er først og fremst å kartlegge norm- og stilklynger (er visse morfologiske valg i noen grad assosiert med en viss syntaktisk stil?), og å belyse den forfatter-individuerende evne som ligger i valgfriheten innenfor norsk skriftspråksnorm. I tillegg antar vi at metoden vil kunne supplere de eksisterende statistiske metodene for forfatterattribusjon, som er basert på mer overfladiske egenskaper ved tekstene. Det synes likevel mer usikkert om den alene vil kunne gi bedre resultater enn de, av grunner som utdypes i konklusjonen (avsnitt 5).

## 2 Åtte grammatiske fenomener

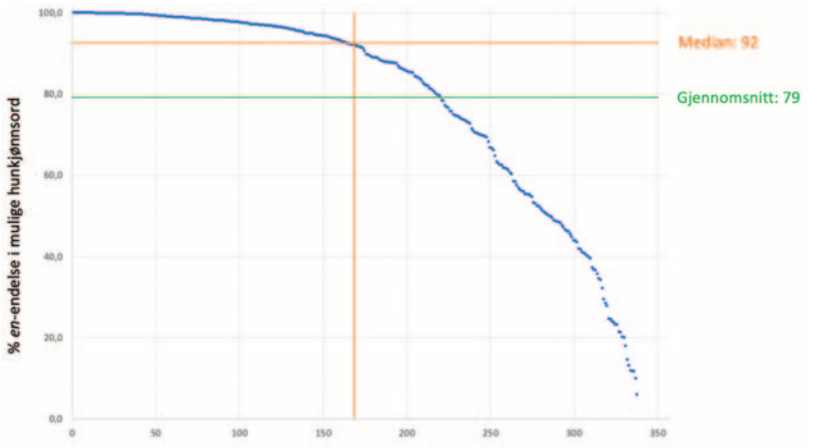
Studien er basert på åtte grammatiske fenomener, tre morfologiske og fem syntaktiske. De tre morfologiske (1–3) er valgt som de antagelig mest sentrale morfologiske valgalternativene i karakteristikken av 'konservativt' vs. 'radikalt' bokmål. De fem syntaktiske (4–8) ansees i noen grad på tilsvarende måte å skille mellom mer og mindre konservativt, formelt språk.

grunner: De er foreløpig skjemmet av en del OCR-skanningfeil som etter hvert blir rettet.

1. -en *vs.* -a *i bestemt form entall av mulige hunkjønnsord;*
2. en *vs.* ei *som ubestemt artikkel ved mulige hunkjønnsord;*
3. -et *vs.* -a *i preteritum av svake verb;*
4. *enkel vs. dobbelt bestemthet;*[9]
5. *foranstilt vs. etterstilt possessiv;*
6. *setningskompleksitet;*
7. være *vs.* ha *som hjelpeverb i perfektum av overgangsverb;*
8. s-*passiv vs.* bli-*passiv.*

Før vi ser etter normklynger av flere av fenomenene, presenterer vi fenomenene enkeltvis og studerer fordelingen av de alternative uttrykkene for dem blant de 338 forfatterne.

*2.1* -en *vs.* -a *i bestemt form entall av mulige hunkjønnsord*
Fig. 1 viser fordelingen av endelsene *-en* og *-a* blant de 338 forfatterne. Forfatterne er ordnet langs *x*-aksen fra dem med mest *-en* til dem med minst *-en*. *y*-aksen viser prosent *-en* i bestemt form entall av mulige hunkjønnsord, basert på de 148 substantivene av denne kategorien som har




*Figur 1.* -en *vs.* -a *i bestemt form entall av mulige hunkjønnsord blant 338 forfattere.*

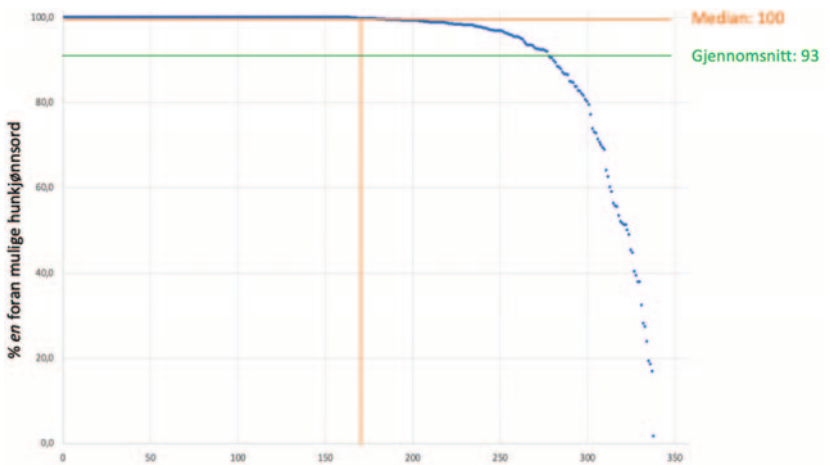
9. Enkel vs. dobbelt bestemthet innebærer morfologisk alternasjon mellom bestemt og ubestemt form, men betraktes likevel som et syntaktisk trekk her fordi alternasjonen er begrenset til en viss syntaktisk konstruksjon – valget står mellom to versjoner av denne. Det er ikke snakk om en generell morfologisk valgfrihet.

mer enn 1000 treff i bestemt form entall i dette materialet (se listen over substantiver i Appendiks, avsnitt 6.2).

Med en median[10] på 92 og et gjennomsnitt på 79 er det en tydelig dominans av valget *-en*. Likevel er ikke dominansen sterkere enn at dette morfologiske valget er blant dem som differensierer best mellom forfattere. Denne differensieringen gjør endelsen i bestemt form entall av mulige hunkjønnsord velegnet som grunnlag for å se etter korrelasjoner med syntaktiske fenomener.

*2.2* en *vs.* ei *som ubestemt artikkel ved mulige hunkjønnsord*

Fig. 2 viser fordelingen av ubestemt artikkelform (determinativform) *en* og *ei* blant forfatterne. Forfatterne er ordnet langs *x*-aksen fra dem med mest *en* til dem med minst *en*. *y*-aksen viser prosent *en* som ubestemt artikkel ved mulige hunkjønnsord, basert på de samme 148 substantivene av denne kategorien som i fig. 1.




*Figur 2.* en *vs.* ei *som ubestemt artikkel ved mulige hunkjønnsord blant 338 forfattere.*

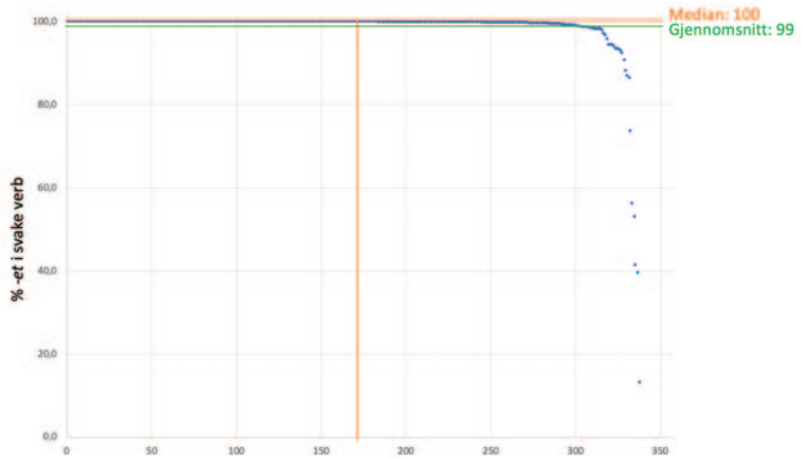
10. Når enhetene er ordnet fra størst til minst, eller omvendt, angir medianen verdien til den midtre enheten (angitt ved en loddrett linje i diagrammene), og dermed den «typiske» verdien for variabelen. Gjennomsnittet kan avvike fra den når få enheter i en ende av skalaen har særlig høye eller lave verdier.

Med en median på 100 og et gjennomsnitt på 93 er dominansen av *en* som ubestemt artikkel ved mulige hunkjønnsord betydelig sterkere enn dominansen av endelsen *-en* i bestemt form entall av de samme substantivene.[11] Den ubestemte artikkelformen differensierer derfor mellom færre forfattere.

*2.3* -et *vs.* -a *i preteritum av svake verb*

Endelsen *-et* dominerer så sterkt over *-a* i preteritum[12] av svake verb av første klasse at dette valget ikke differensierer mellom mange forfattere hvis vi tar alle verbene i betraktning. Den undersøkte delen av trebanken inneholder 1485 ulike verb av denne kategorien. For å oppnå en marginalt bedre differensiering av forfatterne begrenser vi oss til de 196 verbene som har lavest prosent *-et*. (Se listen over de 196 verbene i Appendiks, avsnitt 6.3.)

Fig. 3 viser hvordan endelsene *-et* og *-a* er fordelt blant de 338 forfatterne, for de 196 verbene med lavest prosent *-et* globalt. Også for disse verbene blir differensieringen av forfattere svært begrenset, men det kan likevel være av interesse å sammenholde dette trekket med andre trekk.



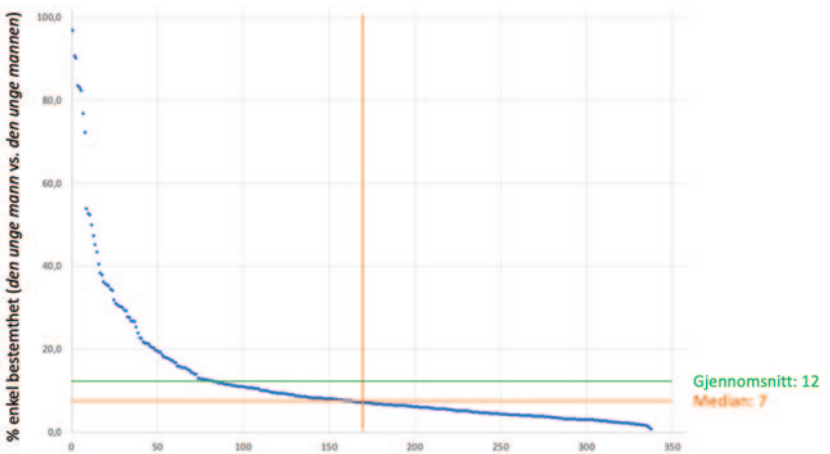

*Figur 3.* -et *vs.* -a *som preteritumsendelse blant 338 forfattere i de 196 verbene som har minst* -et *globalt.*

11. Dette er et kjent forhold, også fra dialekter som for eksempel oslomål, se Lødrup (2011).
12. Supinum viser den samme motsetningen, men denne studien begrenser seg til å telle preteritumsformer ettersom fordelingen i supinum etter alt å dømme er den samme.

*2.4 Enkel vs. dobbelt bestemthet*

Et eksempel på enkel bestemthet er frasen *den unge mann*, mens dobbelt bestemthet illustreres av frasen *den unge mannen*, med både demonstrativ og bestemt form av substantivet. Enkel bestemthet tilskrives vanligvis en mer formell og konservativ stil enn dobbelt bestemthet. De to alternativene er også fordelt ulikt ved ulike substantiver, noe vi ikke skal gå nærmere inn på her; vi skal bare se på hvordan de to typene fordeler seg på forfatterne, på tvers av ulike substantiver.

Fig. 4 viser fordelingen av enkel og dobbelt bestemthet blant forfatterne. Forfatterne er ordnet langs $x$-aksen fra dem med prosentvis mest til dem med prosentvis minst enkel bestemthet, og $y$-aksen viser prosentverdien for enkel bestemthet.




*Figur 4. Enkel vs. dobbelt bestemthet blant 338 forfattere.*

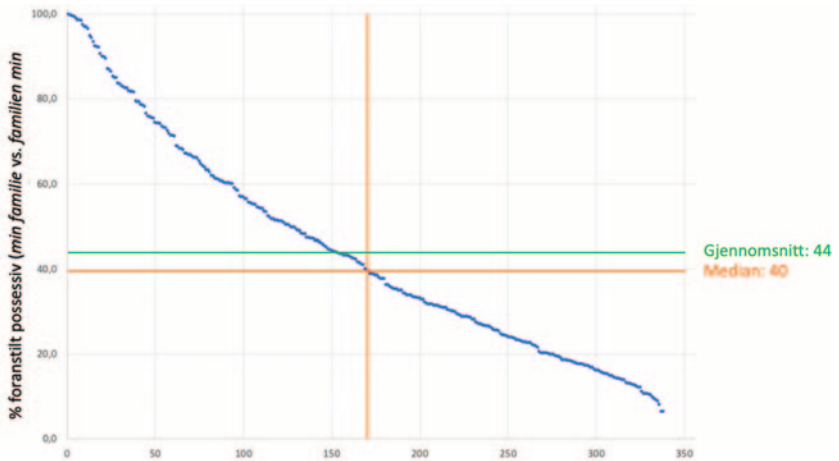
Fig. 4 viser at dobbelt bestemthet dominerer sterkt blant forfatterne, med en median på bare 7 % og et gjennomsnitt på 12 % for alternativet enkel bestemthet. Det høyere gjennomsnittet enn medianen skyldes særlig de få delvis sterkt avvikende forfatterne med over 40 % enkel bestemthet til venstre i grafen.

*2.5 Foranstilt vs. etterstilt possessiv*

Eksempler på foranstilt possessiv er fraser som *min familie*, *hans gamle sykkel*, mens de tilsvarende frasene med etterstilt possessiv er *familien*

*min, den gamle sykkelen hans*, der determinativen 'den' settes inn når substantivet har et attributivt adjektiv foran. I likhet med enkel bestemthet ansees foranstilt possessiv vanligvis som en mer formell og konservativ uttrykksmåte enn alternativet. Også ved foran- vs. etterstilt possessiv er fordelingen mellom de to ulik ved ulike substantiver, men her skal vi bare betrakte hvordan de to typene fordeler seg på forfatterne, uavhengig av substantiv.[13]

Fig. 5 viser fordelingen av foran- og etterstilt possessiv blant de 338 forfatterne, med forfatterne ordnet langs *x*-aksen med dem med høyest andel foranstilt possessiv fra venstre. *y*-aksen viser prosentandel foranstilt possessiv.




*Figur 5. Foranstilt vs. etterstilt possessiv blant 338 forfattere.*

Med en median på 40 og et gjennomsnitt på 44 ser vi at foranstilt possessiv er betydelig vanligere i tekstene, sammenlignet med alternativet, enn enkel bestemthet er sammenlignet med sitt alternativ. Det er et tilnærmet jevnt fall blant forfatterne fra nesten 100 til drøyt 6 % foranstilt

13. I trebank-søket utelukker vi eksempler med determinativen 'egen' (*hans egen sønn*), ettersom de ikke har et alternativ med etterstilt possessiv (*den egne sønnen hans* får en annen betydning av 'egen', tilsvarende den i *hans egne sønn*, der *egne* er et adjektiv).

possessiv. Dette trekket er derfor mer velegnet til å differensiere mellom forfatterne enn enkel bestemthet er.

*2.6 Setningskompleksitet*

Setningskompleksitet kan defineres på ulike måter. Vi vil definere målet for en setnings kompleksitet som antallet leddsetninger den inneholder, pluss antallet dominansforhold mellom leddsetningene. Det betyr at kompleksiteten øker jo dypere et gitt antall leddsetninger er innføyet i hverandre. Vi illustrerer med eksempler fra Dag Solstad, der leddsetninger på nivå 1 er gjengitt i blått, leddsetninger på nivå 2, altså innføyet i setning på nivå 1, er gjengitt i grønt, og leddsetninger på nivå 3, altså innføyet i setninger på nivå 2, er gjengitt i rødt:
En setning med tre leddsetninger på samme nivå har kompleksiteten 3:

(1) *Da du hørte at Merete var død, tenkte du da også at det var forferdelig, og at du hadde medynk med meg?* (*T. Singer*)

Hvis én av de tre setningene er innføyet i en annen, blir kompleksiteten 4:

(2) *Dette er ikke forfatteren som underbygger sin egen myte, men forfatteren som endelig er blitt så gammel at han åpent kan snakke om sin strategi.* (*14 artikler på 12 år*)

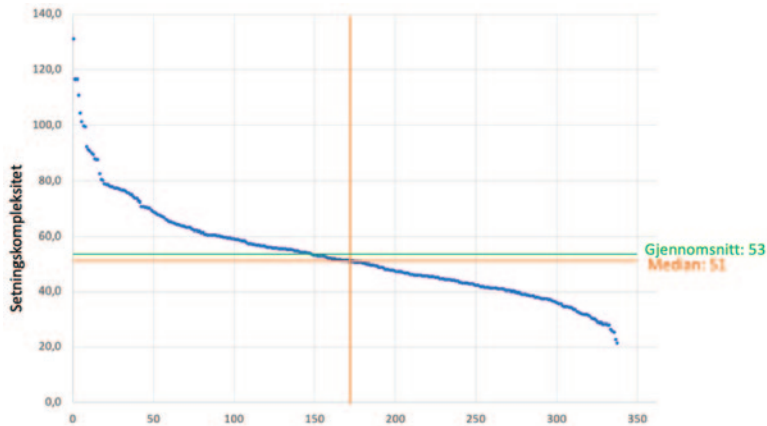
Hvis også den andre er innføyet i den tredje, blir kompleksiteten 6:

(3) *Men hvis disse meningene og disse saksopplysningene blir oppfattet isolert fra det faktum at det er meninger og saksopplysninger som forekommer i en roman, blir ganske mye av vitsen borte.* (*14 artikler på 12 år*)

Kompleksiteten i (3) blir 6 fordi vi i tillegg til de tre leddsetningene har tre dominansforhold: Den blå dominerer (inneholder) den grønne, den grønne dominerer den røde, og den blå dominerer indirekte den røde.

Kompleksiteten i en hel tekst defineres som dens gjennomsnittlige setningskompleksitet ganget med hundre, for å oppnå en størrelsesorden på målet som gjør det mer sammenlignbart med andre kompleksitetsmål som for eksempel LIX/LIKS ('lesbarhetsindeks', se Björnsson 1968). Kompleksiteten i Solstads seks tekster sett under ett er 75,6.

Fig. 6 viser kompleksitetsverdiene i tekstene til de 338 forfatterne, med kompleksitetsverdiene langs *y*-aksen. Kurven viser en tendens til samling rundt et gjennomsnitt på ca. 50, men med et mindre antall klare avvikere fra dette i begge retninger.

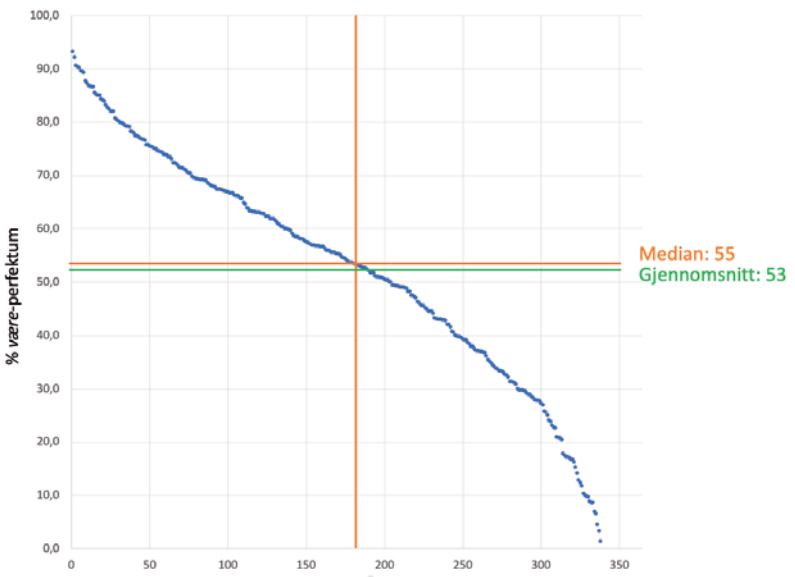



*Figur 6. Setningskompleksitet hos 338 forfattere.*

*2.7* være *vs.* ha *som hjelpeverb i perfektum av overgangsverb*
Eksempler på disse alternative hjelpeverbene er *han var gått hjem* vs. *han hadde gått hjem* og *hun er blitt advart* vs. *hun har blitt advart*. Varianten *være* var mer enerådende i skriftlig bokmål/riksmål tidligere. *ha*-varianten har det sterkeste talemålsgrunnlaget i østnorsk, mens *være* står sterkere i vest- og nordnorsk.

Søk i trebanken etter eksempler på *være* som perfektumshjelpeverb (*den er forsvunnet*) er problematisk på grunn av homonymien med *være* som passiv-hjelpeverb (*den er stjålet*) ved verb som både kan opptre som intransitive overgangsverb og som transitive verb som kan stå i passiv, for eksempel *flytte*. En parser vil ikke ha grunnlag for å avgjøre om *de er flyttet* skal tolkes som 'de har funnet et nytt bosted' (perfektum, = *de har flyttet*), eller som 'noen har flyttet dem' (passiv, = *de er blitt flyttet*), og vil finne begge lesningene, mens det blir noe tilfeldig hvilken av dem den statistiske disambigueringen så velger. Resultatet er at analysene i trebanken i slike tilfeller ikke er pålitelige. For å unngå dette problemet er den gruppen av verb det søkes på, blitt redusert. I trebanken finnes det analyser med *være*-perfektum av ca. 380 verb. Fra denne listen av verb

fjerner vi dem som også kan opptre som transitive verb og som faktisk forekommer i entydig passiv i trebanken. Av de verbene som da gjenstår, fjerner vi også dem som har færre enn 10 forekomster av *være*-perfektum i trebanken. Vi står da igjen med 109 verb (se listen i Appendiks, avsnitt 6.4). Disse verbene ligger til grunn for undersøkelsen av fordelingen av *ha*- og *være*-perfektum blant de 338 forfatterne, vist i fig. 7.



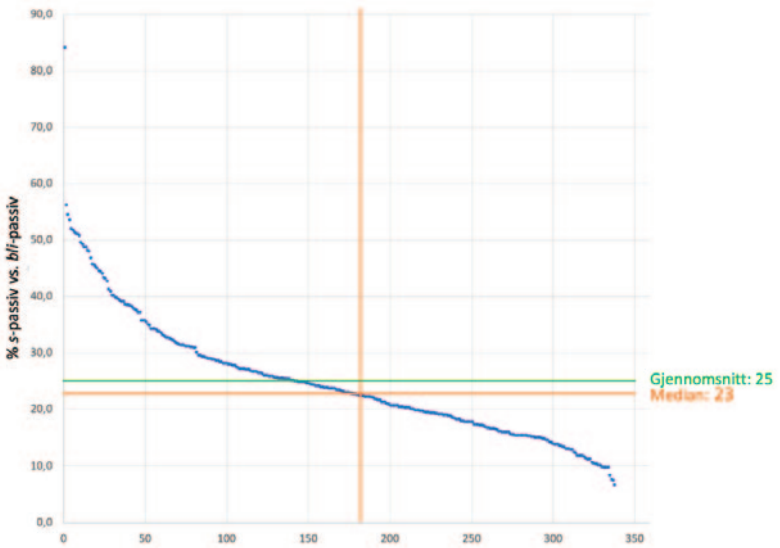

*Figur 7. Fordelingen av* være *og* ha *som hjelpeverb i perfektum av visse verb blant 338 forfattere.*

Fordelingen av *være* og *ha* i fig. 7 antyder ingen tendens i retning av noen binær inndeling av forfattere i *være*-brukere og *ha*-brukere. Det er snarere en svak motsatt tendens i retning av blanding av de to alternativene hos den enkelte.

*2.8* s-*passiv vs.* bli-*passiv*

Eksempler på *s*-passiv (morfologisk passiv) er *referatet skrives av sekretæren, svaret må godkjennes av lederen*, eksempler på *bli*-passiv (perifrastisk passiv) er *referatet blir skrevet av sekretæren, svaret må bli godkjent av lederen*. *s*-passiv assosieres ofte med en mer formell og konservativ stil enn *bli*-passiv, men valget mellom de to kan også være semantisk motivert.

Fig. 8 viser prosentandelen av *s*-passiv regnet av summen av *s*- og *bli*-passiv, fordelt over de 338 forfatterne. Grafen viser altså ikke prosent generell passivbruk sammenlignet med aktiv, men viser fordelingen av de to passivtypene innenfor rammen av passivkonstruksjoner.




*Figur 8. Fordelingen av* s-*passiv og* bli-*passiv blant 338 forfattere.*

Fig. 8 viser at *s*-passiv svært sjelden utgjør mer enn 50 % av passivforekomstene hos en forfatter; *bli*-passiv dominerer sterkt.

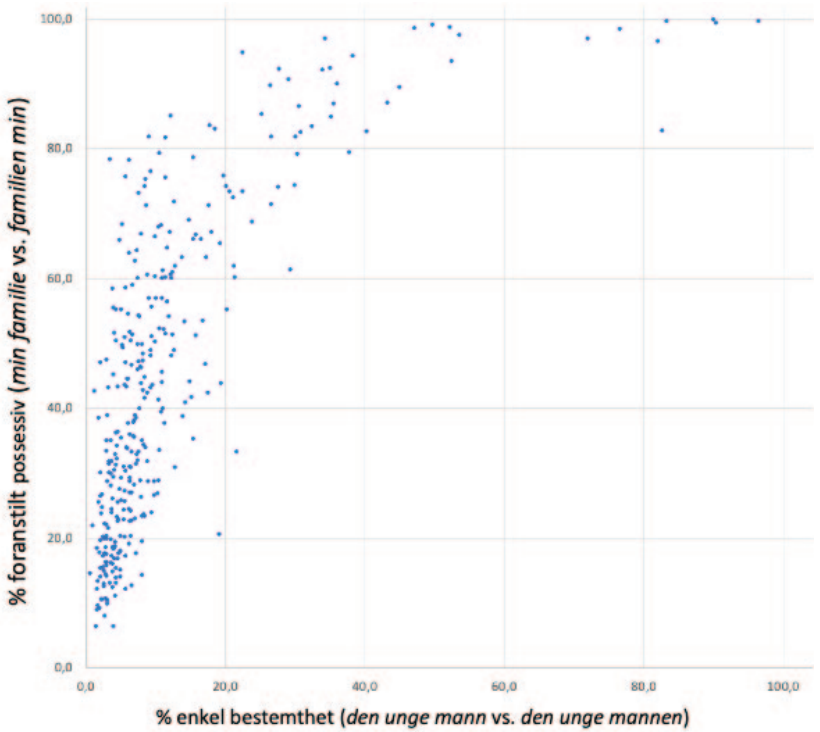
## 3 Normklynger og enkeltforfattere

Avsnitt 2 viste hvordan de alternative uttrykkene for åtte grammatiske fenomener fordeler seg blant de 338 forfatterne. I dette avsnittet skal vi betrakte noen av disse fenomenene parvis og undersøke i hvilken grad uttrykkene for dem samvarierer og slik grupperer forfatterne og deres tekster i klynger. Parallelt med dette skal vi se nærmere på én av forfatterne og undersøke i hvilken grad forfatterens plassering i forhold til normklyngene bidrar til å gi en unik karakteristikk av hans språk, tilstrekkelig til å gjøre ham til den mest sannsynlige forfatter av sine tekster blant de 338 forfatterne. I avsnitt 4 blir denne undersøkelsen utvidet med

ytterligere ni forfattere og deres plassering i forhold til hverandre og de 329 øvrige forfatterne, og de åtte egenskapene betraktes der samlet og ikke bare parvis.

*3.1 Enkel vs. dobbelt bestemthet og foran- vs. etterstilt possessiv*

Det første klynge-eksempelet sammenholder de to syntaktiske fenomenene enkel vs. dobbelt bestemthet og foran- vs. etterstilt possessiv, som vanligvis tillegges lignende stilistiske egenskaper. Det gir en forventning om likhet i måten de fordeler seg mellom forfatterne på.




*Figur 9. Enkel vs. dobbelt bestemthet og foran- vs. etterstilt possessiv fordelt på 338 forfattere.*

Fig. 9 viser forfatternes fordeling på de to egenskapene enkel vs. dobbelt bestemthet ($x$-aksen, med økende prosentsats enkel bestemthet mot høyre) og foran- vs. etterstilt possessiv ($y$-aksen, med økende prosentsats foranstilt possessiv mot toppen). Hvert punkt er en forfatter.

Korrelasjonen mellom de to egenskapene er tydelig og kan tallfestes til 0,71 etter Pearsons korrelasjonskoeffisient, som er et tall mellom +1 og -1 der +1 er perfekt positiv korrelasjon, -1 er perfekt negativ korrelasjon, og 0 er ingen korrelasjon. Det er altså en ganske sterk tendens til at en forfatter vil bruke en lignende proporsjon enkel bestemthet som foranstilt possessiv. Den tetteste klyngen i diagrammet er nede til venstre, med forfattere som overveiende har dobbelt bestemthet og etterstilt possessiv. Etter hvert som vi beveger oss mot høyre med stadig større andel enkel bestemthet, må vi gradvis lenger opp mot en større andel foranstilt possessiv for å finne forfattere. Det er altså, som ventet, en tendens til at mye enkel bestemthet impliserer mye foranstilt possessiv. Men implikasjonen går ikke like tydelig i motsatt retning – vi finner forfattere med under 10 % enkel bestemthet og opp til 80 % foranstilt possessiv. Det er altså vanligere å 'avvike' ved å ha foranstilt possessiv og likevel dobbelt bestemthet enn det omvendte etterstilt possessiv sammen med enkel bestemthet.

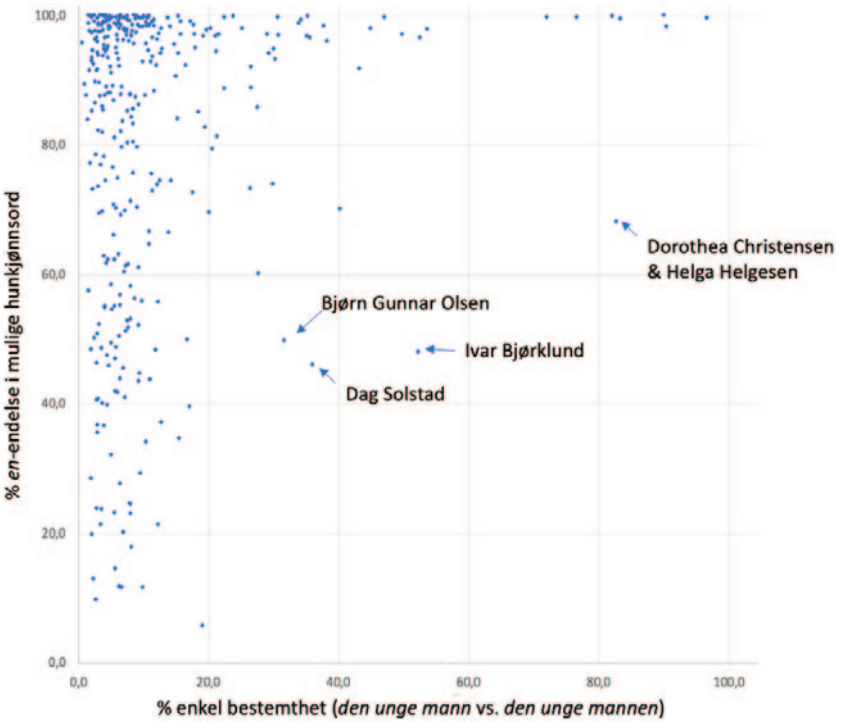
### *3.2* -en *vs.* -a *i mulige hunkjønnsord og enkel vs. dobbelt bestemthet*

#### 3.2.1 Normklynge med avvikere

Det neste paret sammenholder et morfologisk og et syntaktisk fenomen: valget mellom *-en* og *-a* i mulige hunkjønnsord og enkel vs. dobbelt bestemthet.

Fig. 10 viser fordelingen av de 338 forfatterne på egenskapene enkel vs. dobbelt bestemthet (*x*-aksen, med økende prosentsats enkel bestemthet mot høyre) og *-en* vs. *-a* i mulige hunkjønnsord (*y*-aksen, med økende prosentsats *-en* mot toppen). Noen forfatternavn er satt inn i diagrammet; de vil bli kommentert i det som følger.

Den tetteste klyngen av forfattere finnes øverst til venstre, med kombinasjonen *-en*-endelse og dobbelt bestemthet. Den sterke dominansen av *-en* uavhengig av syntaks gjør korrelasjonen mellom de to egenskapene lavere enn den i fig. 9, med en korrelasjonskoeffisient på 0,17, altså en svak positiv korrelasjon. Diagrammet viser likevel sammenhenger. Jo mer ut til høyre med en økende andel enkel bestemthet en forfatter ligger, desto høyere opp mot flere *-en*-endelser har forfatteren en tendens til å ligge. Det indikerer en implikasjon fra mer enkel bestemthet til mer *-en* i mulige hunkjønnsord. Men implikasjonen er ensidig, ettersom en høy andel *-en* ikke impliserer en høy andel enkel bestemthet, men er




*Figur 10.* -en *vs.* -a *i mulige hunkjønnsord og enkel vs. dobbelt bestemthet fordelt på 338 forfattere.*

høyst forenlig med mye dobbelt bestemthet, slik den tette klyngen øverst til venstre viser.

I forbindelse med spørsmålet om forfatterattribusjon skal vi se nærmere på avvikerne fra det mønster at relativt mye enkel bestemthet impliserer relativt mye *-en*-endelse. Avvikerne er de forfatterne som opptrer i den tynt befolkede sentrale delen av diagrammet, der fire er identifisert med navn. Vi skal ta utgangspunkt i *Dag Solstad,* som er representert i trebanken med seks titler:

1. *3 essays*
2. *14 artikler på 12 år*
3. *Ellevte roman, bok atten*
4. *Genanse og verdighet*

5. *Professor Andersens natt*
6. *T. Singer*

Solstads kombinasjon av 36 % enkel bestemthet sammen med 54 % *-a* er atypisk. Stilen hans illustreres av eksempelsetningene i (5); vi har valgt ut noen setninger der hver setning illustrerer begge fenomener:

(5) Eksempler fra Dag Solstad:
- *Å, sola gjennom de fylkeskommunale gardinene i vinduet på dette legekontor på Kongsberg Sykehus!*
- *Han lot sønnen inspisere badet, før han til slutt åpnet døra, på vidt gap, inn til det rom han hadde innredet for Peter.*
- *Og at utsondringene fra sønnens parfymer, bodylotion, after-shave, stick-deodorant, sjampo etc. etc. hang i lufta når han omsider kunne tre inn dit, og overdøvet de mer primitive lukter fra sønnens indre, […]*
- *Egentlig uttrykker Thiis-Evensens historiesyn begeistring over at også historia er i stand til å uttrykke rike menneskers uavhengighet og suverene livsførelse, den suverenitet som kan uttrykkes i det triste og eldgamle faktum:*
- *Saka er jo at når det kulturbærende sjikt oppløses, blir vi alle konsumenter, enten vi vil det eller ikke, […]*

De relevante forfatterne for spørsmålet om forfatterattribusjon av Solstads tekster er de to nærmeste naboene i diagrammet, Ivar Bjørklund og Bjørn Gunnar Olsen.[14]

*Ivar Bjørklund*, en historiker fra Finnmark, er representert i trebanken med verket *Fjordfolket i Kvænangen: fra samisk samfunn til norsk utkant 1550–1980*. Der finner vi eksempler som *Kassa var tom, den eneste syssel-*

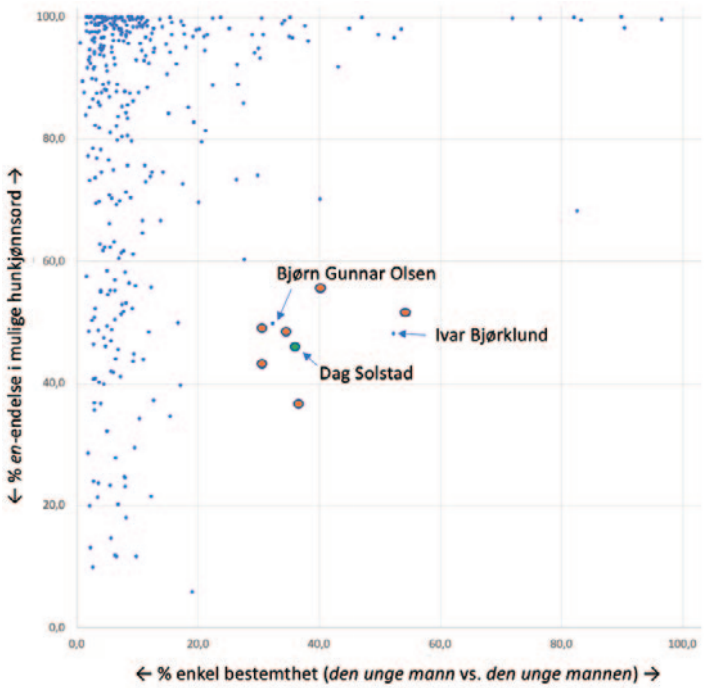
14. Et enda mer avvikende forfatterpar, *Dorothea Christensen og Helga Helgesen*, opptrer lenger ute til høyre i diagrammet, med 83 % enkel bestemthet kombinert med 32 % *-a*. Dette er imidlertid en lite representativ tekst. Det er 9. utgave, fra 1941, av *Husstell for fortsettelsesskoler, realskoler, ungdomsskoler, lærerskoler og skolekjøkkenkurser for voksne*, som først utkom i 1891 på den tids norsk-dansk, og som senere gjennomgikk en serie innholdsmessige og språklige revisjoner utført av andre enn forfatterne. 1941-utgaven (trebankens tekst) har undergått språklige revisjoner i tidens ånd, og de har rammet morfologien mer konsekvent enn syntaksen. Derfor har boken eksempler som: *La fettet bli lysebrunt, trekk panna til side og legg den melete sild i en tett ring i panna med halen inn-mot midten.* Slik halvveis 'modernisering' gir på den måten lett et språk som ingen spontant skriver.

*setting som foregikk var litt nødsarbeide på veien, finansiert gjennom offentlige tilskudd*; *Vel fremme i fjorden kunne han så ikke gjøre annet enn å føre opp Reder på lista over de skatteytere som var rømt siden forrige år.*

*Bjørn Gunnar Olsen,* en journalist og forfatter fra Halden, er representert med 12 titler. Eksempler: *En indolent ungpike herjet og herset med dette åndsmenneske*; *Skulle jeg gi meg til kjenne og ødelegge det opphøyde øyeblikk?*; *Han skriver mye riktig om kjærligheta*; *Ho levde med alle sine krefter, midt i virkeligheta.* De to typene finnes til dels i ulike tekster.

3.2.2 Attribusjon av Solstads tekster på grunnlag av to trekk

Kunne hver av Solstads tekster ha blitt attribuert unikt til Solstad blant de 338 forfatterne bare på grunnlag av egenskapene i 3.2: *-en* vs. *-a* i mulige hunkjønnsord og enkel vs. dobbelt bestemthet? For å vurdere det må vi undersøke hvor spredt tekstene plasserer seg i diagrammet i fig. 10. Det er vist i fig. 11.

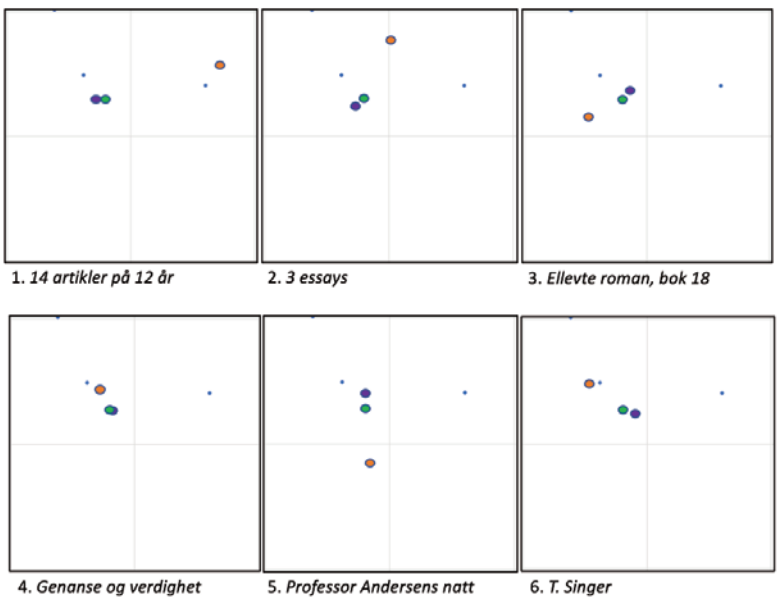



*Figur 11. Solstads seks tekster (røde punkter) plassert i diagrammet for* -en *vs.* -a *i mulige hunkjønnsord og enkel vs. dobbelt bestemthet.*

I fig. 11 er Solstads tekster markert med rødt, mens Solstad som helhet – det vil si tekstene samlet – er markert med grønt. Som det fremgår, ligger tekstene ganske spredt, selv om alle kan sies å avvike fra det mønsteret vi ellers ser i dette diagrammet.

Hvis vi anser hver av de seks tekstene etter tur som skrevet av ukjent forfatter, er vårt sammenligningsgrunnlag for å finne den mest sannsynlige forfatteren den samlede verdien av *resten* av Solstads tekster, i tillegg til tekstene av de nærmeste naboene Bjørklund og Olsen. Fig. 12 viser denne analysen for hver av Solstads seks tekster etter tur. Olsen og Bjørklund er som før representert med punkter, for øvrig er kodingen slik:

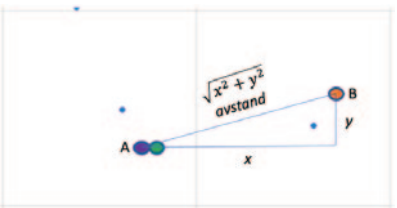
- Teksten under vurdering (rød)
- Resten av Solstads tekster (lilla)
- Alle Solstads tekster (grønn)




*Figur 12. Hver av Solstads seks tekster (røde) etter tur ansett som av ukjent forfatter og sammenlignet med resten av Solstads tekster (lilla) og konkurrentene Bjørklund og Olsen (punkter).*

Spørsmålet er da hvilket punkt den røde teksten ligger nærmest: den lilla som representerer resten av Solstads tekster eller ett av punktene Bjørklund og Olsen?

Avstanden mellom to punkter i de todimensjonale diagrammene regnes ut ved hjelp av Pythagoras. For eksempel skal vi i fig. 12 nr. 1 regne ut avstanden mellom Solstads tekst (rød) og resten av Solstand tekster (lilla). I fig. 13 er $x$ avstanden mellom tekstene A og B langs den vannrette aksen (altså ulikheten mellom A og B i prosentverdi for enkel bestemthet), og $y$ er avstanden mellom A og B langs den loddrette aksen (altså ulikheten mellom A og B i prosentverdi for *-en* i mulige hunkjønnsord). Disse utgjør en rettvinklet trekant sammen med avstanden mellom A og B i det todimensjonale diagrammet, som blir hypotenusen og dermed har en lengde som er kvadratroten av $x^2 + y^2$.




*Figur 13. Avstanden mellom A og B som hypotenusen i en rettvinklet trekant.*

For et par av to egenskaper målt i prosent (0–100) som her, er det tale om en differanse på en skala fra 0 til ca. 140.

- **Tekst 1**, en sakprosatekst, adskiller seg fra Solstad-snittet med et markant høyere innslag av enkel bestemthet. Som man ville forvente, er det Solstads to sakprosatekster som skiller seg ut den veien. Tekst 1 ligger nærmest sakprosaforfatteren Bjørklund (avstand 4,0); avstanden til resten av Solstads tekster er 20,6.
- **Tekst 2**, også en sakprosatekst, adskiller seg fra snittet med flere *en*-endelser og et større innslag av enkel bestemthet. Den ligger nærmere Olsens tekster (avstand 10,4) enn resten av Solstads (avstand 12,0), til tross for den større forskjellen i enkel bestemthet.
- **Tekst 3** adskiller seg fra snittet av Solstads tekster med litt flere *a*-endelser i hunkjønn og færre tilfeller av enkel bestemthet. Den ligger litt nærmere Olsens tekster (avstand 7,1) enn resten av Solstads tekster (avstand 8,1).

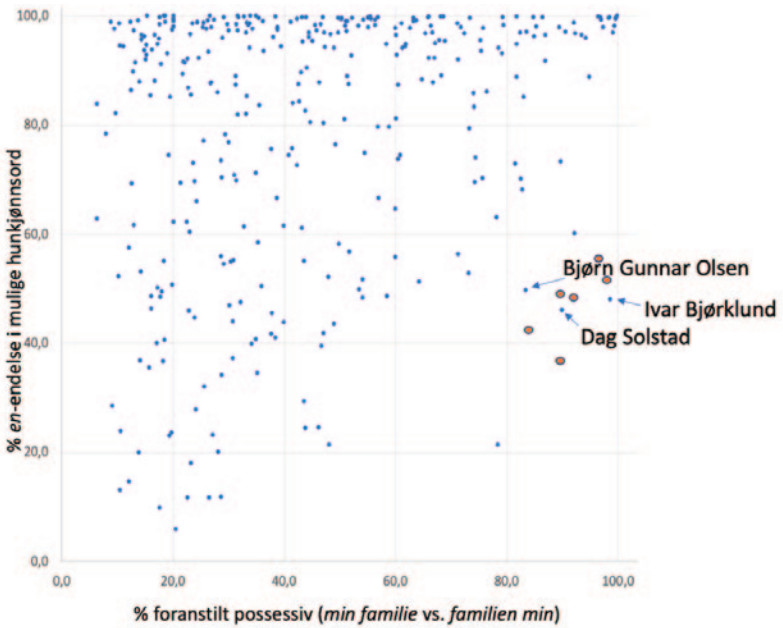
- **Tekst 4** har litt mer *-en* og litt mindre enkel bestemthet enn Solstad-snittet og ligger marginalt nærmere Olsen (avstand 3,1) enn resten av Solstads tekster (avstand 3,3).
- **Tekst 5** har tydelig flere *a*-endelser enn Solstads snitt og havner et stykke unna i diagrammet, men er likevel den eneste teksten som er nærmere Solstads øvrige tekster (avstand 11,1) enn den er både Bjørklund (avstand 19,1) og Olsen (avstand 13,9).
- **Tekst 6** adskiller seg fra snittet på lignende måte som tekst 4, med mer *-en* og mindre enkel bestemthet, og ligger klart nærmere Olsen (avstand 1,0) enn den gjør resten av Solstads tekster (avstand 8,4).

Vi ser at de to trekkene *-en* vs. *-a* i mulige hunkjønnsord og enkel vs. dobbelt bestemthet er tilstrekkelige til å redusere antallet sannsynlige forfattere av Solstads seks tekster fra 338 til 3, men Solstad identifiseres som den mest sannsynlige av de tre ved bare én av de seks tekstene. Spørsmålet er nå hvilke ytterligere trekk som må til for å identifisere Solstad unikt som den mest sannsynlige forfatteren – hvis det er mulig. Vi skal først se på kombinasjonen av *-en* vs. *-a* i mulige hunkjønnsord og foranstilt vs. etterstilt possessiv.

*3.3* -en *vs.* -a *i mulige hunkjønnsord og foran- vs. etterstilt possessiv*

Figur 14 viser fordelingen av de 338 forfatterne på egenskapene *-en* vs. *-a* i mulige hunkjønnsord og foran- vs. etterstilt possessiv, med Bjørklund, Olsen og Solstad markert. Solstads seks tekster er satt inn, markert med rødt.

Som det fremgår av fig. 5 og fig. 6 er foranstilt possessiv betydelig vanligere enn enkel bestemthet i korpus. Langt flere forfattere sprer seg derfor mot høyre og mot høyere andel foranstilt possessiv i fig. 14 enn antallet som sprer seg mot høyre og mot mer enkel bestemthet i fig. 10. Vi får også en høyere korrelasjon mellom de to aksene i fig. 14 – den er på 0,26, sammenlignet med 0,17 i fig. 10. Også i fig. 14 er det nedre høyre del av diagrammet som er tynnest befolket; få forfattere kombinerer mange *a*-endelser med mye foranstilt possessiv. Etter hvert som vi beveger oss fra venstre kant mot høyre og mer foranstilt possessiv, må vi gradvis høyere opp mot mer *-en* for å finne forfattere. Som ved enkel bestemthet er ikke implikasjonen like sterk den andre veien – mange




*Figur 14.* -en *vs.* -a *i mulige hunkjønnsord og foran- vs. etterstilt possessiv fordelt på 338 forfattere. Solstads seks tekster er inkludert og markert med rødt.*

forfattere bruker mye *-en* og holder seg likevel mest til etterstilt possessiv.[15]

Også i fig. 14 skiller Bjørklund, Olsen og Solstad seg ut fra klyngen, og de holder sammen som før – ikke bare langs *y*-aksen, som er den samme, men også langs *x*-aksen, alle med over 80 % foranstilt possessiv.

(6) Eksempler fra Dag Solstad:

- *Etterpå ble døra til hennes rom åpnet*
- *Det er ikke jeg som skaper framtida, selv om mitt ambisjonsnivå forøvrig er skyhøyt.*
- *Camilla hadde da nettopp begynt på skolen, og mora tok brevet med seg inn på hennes rom, og leste det for henne der.*

15. Den ensomme forfatteren nede til høyre med nesten 80 % *-a* og nesten 80 % foranstilt possessiv er Tron Øgrim.

- *Da klokka nærmet seg tolv, reiste han seg fra sin komfortable lenestol.*
- *Hans unge kone, Mette, satt i stua og ammet deres barn.*

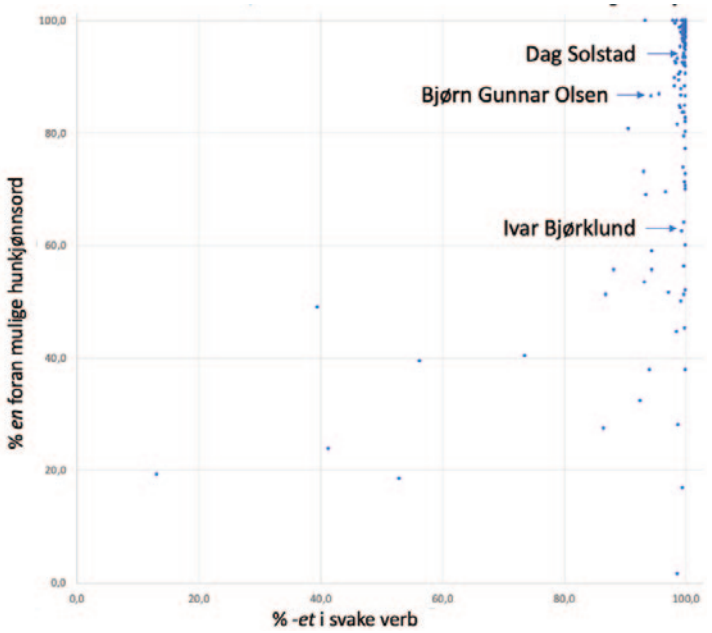
Ivar Bjørklund har eksempler som *Til sist mistet de også sine juridiske rettigheter til jorda i fjorden og med det var samene og deres hjemplasser en del av det dansk-norske rike på lik linje med dets øvrige innbyggere*; *En avklaring av deres innhold vil ha stor betydning for framtida*. Bjørn Gunnar Olsen har blant annet eksemplene: *Døm åpna vinduene med sine fabrikkgustne ansikt, fant fram glass og filmbetær, venta på at klokka skulle bli tre minutter på halv ett, og månen kaste sine skyggær for sola*; *Simon betrakta søstera med sine milde blå øyne*.

Plasseringen av Solstads seks tekster (røde punkter i fig. 14) tyder på at foran- vs. etterstilt possessiv ikke bidrar til å skille klarere mellom de tre forfatterne Solstad, Bjørklund og Olsen (en mer nøyaktig utregning kommer i avsnitt 4). Vi skal se på to ytterligere egenskaper der de alternative mulighetene er langt skjevere fordelt i tekstene: ubestemt artikkel *en* vs. *ei* ved mulige hunkjønnsord og endelsen *-et* vs. *-a* ved svake verb.

### *3.4* en *vs.* ei *som ubestemt artikkel og* -et *vs.* -a *ved svake verb*

Avsnittene 2.2 og 2.3 viste at begge egenskapene *en* vs. *ei* som ubestemt artikkel og (særlig) *-et* vs. *-a* som endelse i svake verb er svært skjevt fordelt i tekstene, med henholdsvis *en* og *-et* som de sterkt dominerende. En kombinasjon av disse to kan vise i hvilken grad valgene blant dem er korrelert. Fig. 15 viser denne kombinasjonen, med prosentandel *-et* i svake verb langs $x$-aksen og prosentandel *en* som ubestemt artikkel i mulige hunkjønnsord langs $y$-aksen. De tre forfatterne Solstad, Bjørklund og Olsen er markert.

Korrelasjonen mellom de to egenskapene er forholdsvis god, med 0,56 som korrelasjonskoeffisient, til tross for den lave frekvensen av *-a*. De seks forfatterne i nedre del mot venstre, som har mindre enn 80 % *-et*, har alle mindre enn 60 % *en*. Den dominerende normklyngen er øverst til høyre med henholdsvis *-et* og *en*, og i dette tilfellet ligger også Solstad i klyngen, med 99,8 % *-et* i svake verb og 92 % *en* som ubestemt artikkel. Olsen har et lite innslag av *-a* (6 %), men ligger fremdeles nær Solstad. Bjørklund skiller seg ut med et sterkere innslag av *ei* som ubestemt artikkel: 37 %. For å vurdere om avstanden er stor nok til å gjøre Bjørklund mindre sannsynlig enn Solstad som forfatter må vi sammenligne tekstenes fordeling av *en* og *ei* enkeltvis med Bjørklunds fordeling. En slik




*Figur 15.* -et *vs.* -a *i svake verb og* en *vs.* ei *som ubestemt artikkel ved mulige hunkjønnsord.*

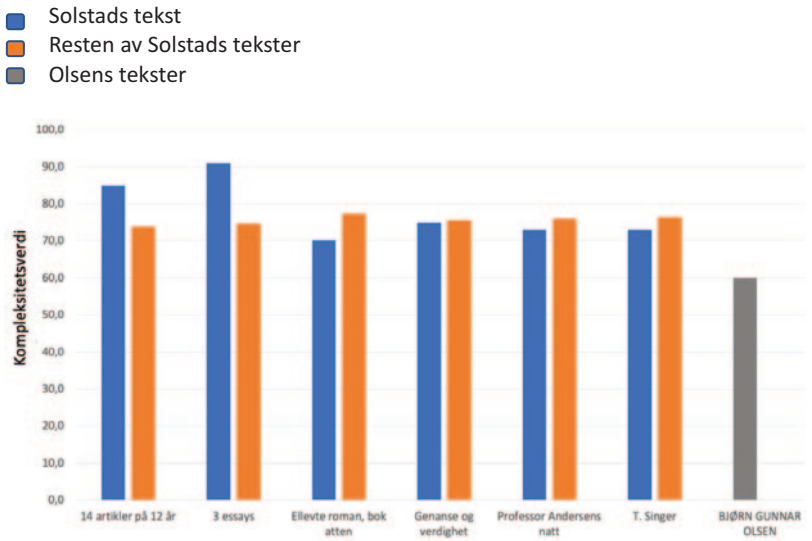
sammenligning viser at alle seks tekster kommer nærmere resten av Solstads tekster enn de kommer Bjørklunds tekst i bruken av *en* vs. *ei* som ubestemt artikkel. *Ellevte roman, bok atten* har den laveste andelen *en*, 80%, men kommer likevel nærmere resten av Solstad-tekstene (94,8 %) enn den kommer Bjørklunds tekster (62,5 %).

Med tilføyelsen av egenskapen *en* vs. *ei* som ubestemt artikkel er Ivar Bjørklunds sjanser i konkurransen om å ha skrevet Solstads tekster dermed redusert. Bjørn Gunnar Olsen gjenstår som en klarere konkurrent. Vi tester egenskapen setningskompleksitet (se avsnitt 2.6) som et mulig grunnlag for å utelukke ham.

*3.5 Setningskompleksitet*

Fig. 6 i avsnitt 2.6 viser kompleksitetsverdiene i tekstene til de 338 forfatterne. Solstad har verdien 75,6 og Olsen verdien 59,9 – begge over median og gjennomsnitt, men likevel med relativt stor avstand mellom dem: Solstad er rangert som nr. 34 blant forfatterne, og Olsen som nr. 91. Kon-

sekvensen av dette for attribusjonen av Solstads seks tekster vises i fig. 16, med følgende fargekoding:



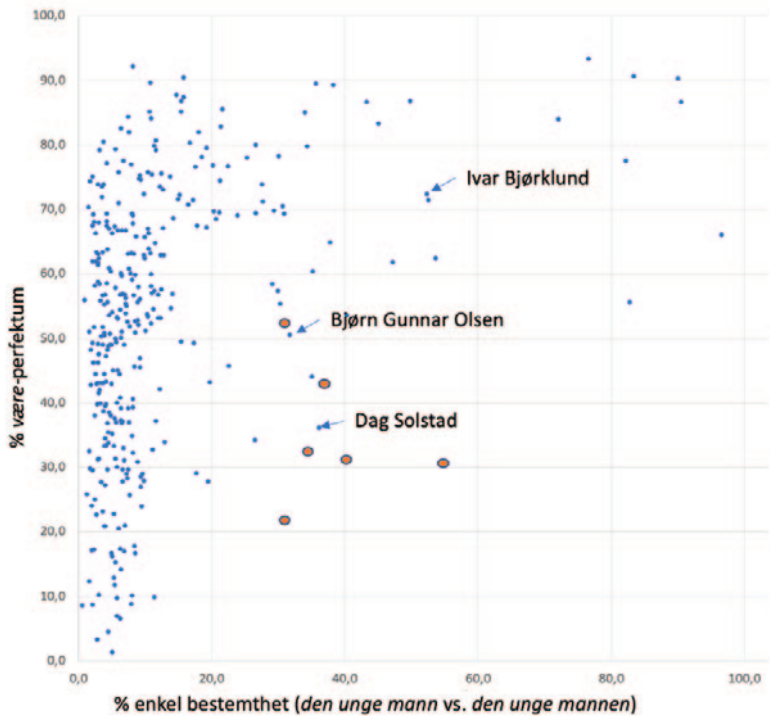

*Figur 16. Setningskompleksiteten i hver av Solstads tekster (blått) sammenlignet med kompleksiteten i resten av Solstads tekster (oransje) og kompleksiteten i Olsens tekster (grått).*

Som man kan vente, skiller de to sakprosatekstene *14 artikler på 12 år* og *3 essays* seg ut som de eneste der kompleksiteten er større enn snittet av Solstads øvrige tekster. Men også teksten med lavest kompleksitet, *Ellevte roman, bok atten*, med en kompleksitet på 70,2, ligger nærmere Solstads øvrige tekster (77,2) enn den gjør Olsens tekster (59,9). Ved å inkludere setningskompleksitet som det fjerde trekket har vi styrket Dag Solstads sjanser i konkurransen om å være den mest sannsynlige forfatter av hver av sine seks tekster blant 338 forfattere. Vi skal se på egenskapene samlet i avsnitt 4.

*3.6* være *vs.* ha *som hjelpeverb i perfektum av overgangsverb og enkel vs. dobbelt bestemthet*

Som nevnt i avsnitt 2.7 er valg av *være* som perfektumshjelpeverb ansett som typisk for noe konservativt språk samtidig med at det har et sterkere talemålsgrunnlag i vest og nord enn det har i øst i Norge. Fig. 17 viser

hvordan dette valget er korrelert med et annet valg som ansees som typisk for konservativt språk, enkel bestemthet. Solstads seks tekster er med i rødt.




*Figur 17. Enkel bestemthet og valg av* være *som hjelpeverb i perfektum av overgangsverb, med Solstads tekster inkludert (røde punkter).*

Korrelasjonen mellom de to aksene er ikke ubetydelig, med en koeffisient på 0,39. Tendensen er synlig i diagrammet, der et sterkere innslag av enkel bestemthet mot høyre henger sammen med plassering av forfatteren i øvre del, med mer *være*-perfektum. Men den tettere klyngen i venstre del av diagrammet viser at *være*-perfektum går godt sammen med dobbelt bestemthet også – tilløpet til implikasjon er også her ensidig.

Plasseringen av de tre navngitte forfatterne og Solstads tekster (rødt) indikerer at også bruk av *være*-perfektum skiller tydelig mellom Solstad og Bjørklund.

## 4 Sammenligning av ti forfatteres tekster

Vi har sett på et utvalg morfologiske og syntaktiske egenskaper og hvordan de er korrelert i klynger hos 338 forfattere, og vi har særlig sett på ett forfattereksempel, Dag Solstad, blant dem som avviker fra en normklynge. Bare to andre forfattere, Ivar Bjørklund og Bjørn Gunnar Olsen, skilte seg ut som konkurrenter ved forfatterattribusjon av Solstads seks tekster da vi betraktet normklyngen som var basert på egenskapene *-en* vs. *-a* i bestemt form entall av mulige hunkjønnsord og enkel bestemthet. Dag Solstad styrker sin stilling som den mest sannsynlige forfatter av hver av de seks tekstene han er representert med, når vi i tillegg tar i betraktning egenskapene *en* vs. *ei* som ubestemt artikkel ved mulige hunkjønnsord og setningskompleksitet. Denne kasusstudien indikerer dermed muligheten for at et lite antall normvalg (inklusive syntaksvalg) i tekstene til en forfatter i utkanten av en eller flere normklynger kan identifisere forfatteren unikt blant flere hundre konkurrenter. Det gjenstår å se hvordan de ulike grammatiske egenskapene slår ut samlet i den endelige kåringen av den mest sannsynlige forfatteren av Solstads tekster. Andre gjenstående spørsmål er i hvilken grad forfattere som i mindre grad avviker fra normklynger, lar seg sirkle inn på grunnlag av normvalgene i sine tekster, og om resultatet for Solstad lar seg replikere hos andre klyngeperifere forfattere. Også andre faktorer enn normklyngeplassering kan tenkes å påvirke muligheten for forfatterattribusjon. Vi tester disse spørsmålene for et utvalg forfattere i dette avsnittet.

### *4.1 Ni tilfeldig valgte forfattere pluss Solstad*

I tillegg til Solstad ble ni av de 338 forfatterne valgt ut tilfeldig med en random-funksjon, med den begrensning at hver forfatter måtte være representert med minst to tekster på minst 1000 analyserte setninger hver, og halvparten av forfatterne med minst fire tekster. Forfatterne og deres tekster, Solstad inkludert, vises i Tabell 1.

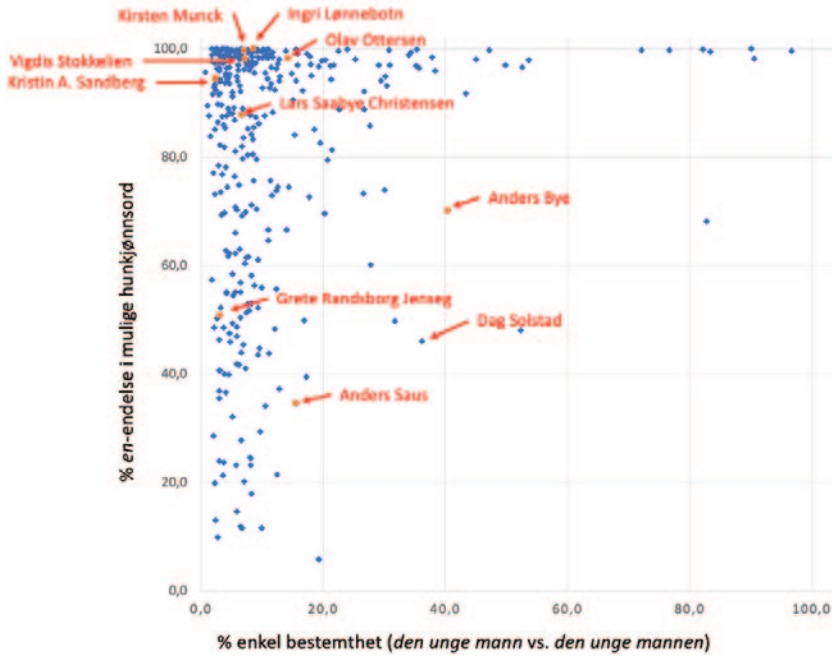
| Forfatter | Tekst | Analyserte setninger |
|---|---|---|
| **Anders Bye** | 1. *En eventyrer bekjenner: Sara - tredje sats* | 5438 |
| | 2. *Hyggepianisten* | 3956 |
| | 3. *Mannen som ville bli menneske: roman* | 5297 |
| | 4. *Rondo: sonaten om Sara-fjerde sats* | 4612 |
| | 5. *Sara og Iversen: chaconne* | 4635 |
| | 6. *Ute* | 3176 |

| Forfatter | Tekst | Analyserte setninger |
|---|---|---|
| **Lars Saabye Christensen** | 1. *Amatøren* | 3472 |
| | 2. *Columbus' ankomst* | 2671 |
| | 3. *Den misunnelige frisøren* | 3216 |
| | 4. *Gutten som ville være en av gutta* | 2269 |
| | 5. *Herman* | 4953 |
| | 6. *Ingens* | 3598 |
| | 7. *Noen som elsker hverandre* | 2720 |
| | 8. *Sneglene* | 3581 |
| **Grete R. Jenseg** | 1. *Kjære ingen!* | 5815 |
| | 2. *Nesten nær deg* | 3976 |
| **Ingri Lønnebotn** | 1. *Alene, mot himmelen* | 3022 |
| | 2. *Den andre* | 4128 |
| | 3. *Hun var en liten kvinne jeg ikke kjente særlig godt* | 4364 |
| **Kirsten Munck** | 1. *Nattfugl* | 5870 |
| | 2. *Sort madonna* | 4763 |
| **Olav Ottersen** | 1. *Den som kommer aller sist* | 4058 |
| | 2. *Fangarmer* | 4266 |
| | 3. *Fri som fuglen* | 3210 |
| | 4. *I hevnerens skygge* | 2107 |
| | 5. *I nød og lyst* | 4764 |
| | 6. *Mistenkt* | 3951 |
| | 7. *Over grensen* | 3877 |
| | 8. *Silkenettet* | 4760 |
| | 9. *Skjebnereisen* | 4342 |
| **Kristin A. Sandberg** | 1. *Marcus og den blå hunden* | 1082 |
| | 2. *Nødutgang* | 3194 |
| | 3. *Usynlig* | 2124 |
| **Anders Saus** | 1. *Broder Daniels hender* | 3082 |
| | 2. *Brødrene på Hundholmen* | 2963 |
| | 3. *Fride Berrføtt* | 3300 |
| | 4. *I skyggen av Bjønnvassbre* | 3346 |
| **Dag Solstad** | 1. *14 artikler på 12 år* | 1609 |
| | 2. *3 essays* | 623 |
| | 3. *Ellevte roman, bok atten* | 2436 |
| | 4. *Genanse og verdighet* | 1417 |
| | 5. *Professor Andersens natt* | 1531 |
| | 6. *T. Singer* | 2720 |
| **Vigdis Stokkelien** | 1. *Båten under solseilet* | 2985 |
| | 2. *Stjerneleden* | 2433 |

*Tabell 1. Ti forfattere og deres 45 tekster i NorGramBank*

Figur 18 viser hvordan de ti forfatterne er plassert i forhold til den normklyngen som vises i figur 10 i avsnitt 3.2 (*-en* vs. *-a* i mulige hunkjønnsord og enkel vs. dobbelt bestemthet).

Som ventelig er ved tilfeldig utvalg, er de fleste forfatterne ganske sentralt plassert i normklyngen, mens tre-fire skiller seg noe ut. Dette er da bare én normklynge, mens det er forfatternes plassering i forhold til




*Figur 18. De ti forfatterne plassert i forhold til normklyngen i avsnitt 3.2.*

alle normklyngene i materialet som er relevant; mer komplette data om dette kommer i avsnitt 4.4.

*4.2 Åtte grammatiske egenskaper i ett rom, og avstandsmål*

Vi ønsker å betrakte hver av de 45 tekstene av de ti forfatterne etter tur som skrevet av ukjent forfatter, og så sammenligne den med alle de 338 forfatterne, der tekstens virkelige forfatter representeres med *de øvrige* av sine tekster, for å se hvilken forfatter teksten kommer nærmest, og hvilket nummer i rekken den virkelige forfatteren får. Sammenligningen baserer seg på de åtte grammatiske egenskapene som presenteres i avsnitt 2.

I avsnitt 3 betraktet vi de grammatiske egenskapene parvis og viste normklynger i todimensjonale diagrammer. I dette avsnittet skal vi betrakte de åtte egenskapene under ett. Siden hver forfatter har en plassering langs hver av de åtte aksene, kan forfatterne ansees som plassert i et åttedimensjonalt rom der hver av de åtte aksene utgjør en dimensjon. Vi må da gi avkall på muligheten for å vise klyngestrukturen grafisk på et todimensjonalt papir, men vi kan til gjengjeld regne ut avstanden mel-

lom to forfattere eller tekster langs alle dimensjonene samlet og få én avstandsverdi for hvert par.

Utregningen av avstanden mellom to tekster *A* og *B* i de todimensjonale diagrammene er forklart i avsnitt 3.2.2 som basert på en rettvinklet trekant. Prinsippet er det samme i flerdimensjonale rom med flere akser, i vårt tilfelle åtte; forskjellen er bare at det blir flere rettvinklede trekanter med en linje *AB* som hypotenus, og dermed flere kateter å summere kvadratene av.[16]

*4.3 Resultater for de ti forfatterne*

I databasen bak diagrammene i denne artikkelen er i utgangspunktet en gitt forfatters verdi for en viss variabel lik verdien for alle forfatterens tekster behandlet som én tekst. For å undersøke hvor nær en gitt tekst kommer sin forfatter, må forfatteren i den sammenheng som nevnt representeres med komplementmengden av sine tekster, altså resten av tekstene etter at den undersøkte teksten er trukket fra. Slike komplementmengder er regnet ut for alle de 45 tekstene til de ti forfatterne og lagt til databasen. Avstandene som oppgis nedenfor mellom en undersøkt tekst og dens forfatter, er følgelig alltid avstanden mellom teksten og den relevante komplementmengden av forfatterens tekster.

For hver tekst A appliseres formelen for utregning av avstand (fotnote 16) til forfatteren (representert ved komplementmengden av tekstene) og til alle de resterende 337 forfatterne, som B-tekster. Dette gir 338 avstander som kan sorteres fra den minste til den største.[17] Forfatteren med den minste avstanden til teksten kåres som vinner. Tabell 2 viser resultatet for hver tekst, der *forf. rang* er den egentlige forfatterens plass i listen av avstander fra den minste til den største, *avst. forf* er avstanden mellom

16. For hver egenskap av de åtte subtraherer vi da *B*s verdi fra *A*s verdi for denne egenskapen og finner dermed avstanden langs den aktuelle aksen, som vi tar kvadratet av. Så summerer vi de åtte kvadrerte avstandene og tar kvadratroten av summen. Hvis $v_1 - v_n$ er egenskaper eller variabler som våre åtte, og *A* og *B* er to tekster, kan utregningen av avstanden mellom *A* og *B* i det *n*-dimensjonale rommet dermed formuleres slik:

$$\sqrt{\sum_{i=1}^{n}\left(v_i(A) - v_i(B)\right)^2}$$

17. Med åtte egenskaper eller dimensjoner, og skalaer for hver egenskap fra 0 til 100, blir den maksimale avstanden vi i prinsippet kan oppnå, ca. 283. I praksis ligger de største avstandene med våre åtte egenskaper og 338 forfattere på omkring 195.

teksten og den egentlige forfatteren, *vinner* er forfatteren med kortest avstand til teksten, og *avst. vinner* er avstanden mellom teksten og vinneren. De tilfellene der den egentlige forfatteren er vinneren og dermed har rang 1, er markert med rødt.

| Forfatter | Tekst | Forf. rang | Avst. forf. | Vinner | Avst. vinner |
|---|---|---|---|---|---|
| **Anders Bye** | 1. *En eventyrer ...* | 1 | 19,8 | **Anders Bye** | 19,8 |
| | 2. *Hyggepianisten* | 8 | 28,2 | Hilde Henriksen Waage | 19,5 |
| | 3. *Mannen ...* | 4 | 26,4 | Kristian Kristiansen | 19,4 |
| | 4. *Rondo ...* | 1 | 8,4 | **Anders Bye** | 8,4 |
| | 5. *Sara og ...* | 1 | 21,5 | **Anders Bye** | 21,5 |
| | 6. *Ute* | 1 | 15,5 | **Anders Bye** | 15,5 |
| **Lars Saabye Christensen** | 1. *Amatøren* | 44 | 27,7 | Erlend Aas | 11,7 |
| | 2. *Columbus ...* | 5 | 29,6 | Arne Berggren | 25,5 |
| | 3. *Den misunnelige* | 3 | 15,4 | Kjersti Wold | 8,3 |
| | 4. *Gutten ...* | 70 | 57,3 | Gry Brenna | 26,7 |
| | 5. *Herman* | 4 | 21,4 | Kjersti Wold | 19,4 |
| | 6. *Ingens* | 1 | 12,1 | **Lars Saabye Christensen** | 12,1 |
| | 7. *Noen ...* | 1 | 14,3 | **Lars Saabye Christensen** | 14,3 |
| | 8. *Sneglene* | 15 | 30,0 | Øivind Loraas | 16,0 |
| **Grete R. Jenseg** | 1. *Kjære ingen!* | 15 | 29,3 | Anne B. Ragde | 17,3 |
| | 2. *Nesten nær deg* | 7 | 29,3 | Else Lien Krogedal | 18,0 |
| **Ingri Lønnebotn** | 1. *Alene ...* | 7 | 32,6 | Solfrid Elgvin Lied | 21,4 |
| | 2. *Den andre* | 43 | 36,8 | Selma Ståhl | 12,7 |
| | 3. *Hun var ...* | 43 | 38,3 | Jostein Gaarder | 18,8 |
| **Kirsten Munck** | 1. *Nattfugl* | 36 | 38,0 | Thorbjørn R. Johansen | 13,2 |
| | 2. *Sort madonna* | 72 | 38,0 | Mona Berg | 8,8 |
| **Olav Ottersen** | 1. *Den som ...* | 2 | 11,2 | Elisabeth Aasen | 7,9 |
| | 2. *Fangarmer* | 4 | 11,4 | Frank Lie | 8,5 |
| | 3. *Fri som fuglen* | 2 | 9,5 | Elisabeth Aasen | 6,8 |
| | 4. *I hevnerens ...* | 2 | 11,5 | Helge Riisøen | 7,8 |
| | 5. *I nød og lyst* | 1 | 13,2 | **Olav Ottersen** | 13,2 |
| | 6. *Mistenkt* | 3 | 11,7 | Finn Jor | 9,3 |
| | 7. *Over grensen* | 1 | 9,2 | **Olav Ottersen** | 9,2 |
| | 8. *Silkenettet* | 1 | 9,1 | **Olav Ottersen** | 9,1 |
| | 9. *Skjebnereisen* | 1 | 8,6 | **Olav Ottersen** | 8,6 |
| **Kristin A. Sandberg** | 1. *Marcus og ...* | 2 | 26,9 | Margaret Aronsen Lykken | 26,5 |
| | 2. *Nødutgang* | 29 | 28,1 | Tove Gravem Smedstad | 11,5 |
| | 3. *Usynlig* | 55 | 29,1 | Unni Wenche Tandberg | 10,6 |
| **Anders Saus** | 1. *Broder ...* | 16 | 27,6 | Roy Jacobsen | 8,9 |
| | 2. *Brødrene på ...* | 1 | 23,3 | **Anders Saus** | 23,3 |
| | 3. *Fride Berrføtt* | 1 | 23,2 | **Anders Saus** | 23,2 |
| | 4. *I skyggen ...* | 1 | 17,4 | **Anders Saus** | 17,4 |
| **Dag Solstad** | 1. *14 artikler ...* | 1 | 23,3 | **Dag Solstad** | 23,3 |
| | 2. *3 essays* | 1 | 31,7 | **Dag Solstad** | 31,7 |
| | 3. *Ellevte roman ...* | 1 | 31,2 | **Dag Solstad** | 31,2 |
| | 4. *Genanse ...* | 1 | 10,2 | **Dag Solstad** | 10,2 |
| | 5. *Professor ...* | 1 | 18,0 | **Dag Solstad** | 18,0 |
| | 6. *T. Singer* | 2 | 24,8 | Bjørn Gunnar Olsen | 19,4 |
| **Vigdis Stokkelien** | 1. *Båten under ...* | 61 | 26,0 | Oddvar Nilsen | 7,4 |
| | 2. *Stjerneleden* | 17 | 26,0 | Svein Arne Laukli | 7,6 |

*Tabell 2. Resultater for ti forfattere og deres 45 tekster.*

Fem av de ti forfatterne oppnår å bli kåret som den mest sannsynlige forfatter av noen av sine tekster: *Anders Bye*, *Lars Saabye Christensen*, *Olav Ottersen*, *Anders Saus* og *Dag Solstad*. Disse fem, eller i hvert fall noen av dem, deler flere egenskaper som skiller dem fra de øvrige forfatterne. Av disse egenskapene er felles kjønn påfallende, men med usikker relevans[18] – men det finnes andre:

1. Som figur 18 viser, er Anders Bye og Anders Saus, i tillegg til Solstad, de to tydeligste avvikerne fra den normklyngen som vises der. Også Lars Saabye Christensen ligger lenger unna sentrum i klyngen enn de fleste andre, men Grete Randsborg Jenseg er likevel mer klyngeperifer enn han. Se avsnitt 4.4 for mer komplette data om dette.

2. Alle fem, Bye, Saabye Christensen, Ottersen, Saus og Solstad, er representert med flere tekster, fra 4 til 9, enn noen av de øvrige forfatterne.

3. Avstandsmålene viser at de atten korrekt attribuerte tekstene, med ett unntak hos Ottersen og to hos Solstad, har kortere avstand til sin egentlige forfatter (avst. forf.) enn de ukorrekt attribuerte tekstene av samme forfatter har.

Egenskap nr. 1 er i overensstemmelse med hypotesen at perifer plassering i normklynger letter forfatterattribusjonen på grunn av mer glissent naboskap i normrommet og dermed færre konkurrenter.

At en forfatter er representert med flere tekster (egenskap 2), kan ha en betydning fordi et høyere antall tekster i komplementmengden til en tekst i noen grad må ventes å jevne ut språklig variasjon mellom forfatterens tekster. Når en forfatter er representert med bare to tekster, er komplementmengden lik den andre teksten alene, og da er risikoen større for stor avstand. Dette innebærer at egenskap nr. 3, tekstens avstand til den egentlige forfatteren, kan være påvirket av egenskap nr. 2: Et høyt antall tekster kan i seg selv føre til mindre avstand mellom en gitt tekst og dens komplementmengde. Siden avstanden mellom en tekst og forfatteren er avstanden mellom teksten og forfatterens øvrige tekster, indikerer den i hvilken grad forfatteren varierer språklig mellom tekstene

18. Riktignok viser materialet for denne artikkelen en tydelig tendens til at avvikerne fra normklynger i overveiende grad er menn, men dette temaet forfølges ikke her.

sine. Jo mer variasjon, desto mindre sjanse for å lykkes med forfatterattribusjonen.

Rangeringen av forfatteren (kolonnen *forf. rang*) i de tilfellene der forfatteren ikke er vinneren, vil være en følge av antallet nære konkurrenter, som igjen vil være påvirket av to faktorer: tettheten i klyngen rundt forfatteren og avstanden mellom teksten og forfatteren. Vi ser at lave rangeringer som nr. 36, 43 og 72 hos *Ingri Lønnebotn* og *Kirsten Munck* er korrelert med særlig høye avstander til forfatteren mellom 36,8 og 38. Den lave plasseringen som nr. 70 ved teksten *Gutten som ville være en av gutta* hos *Lars Saabye Christensen* er korrelert med en avstand på hele 57,3 mellom teksten og resten av Saabye Christensens tekster, noe som tyder på at denne teksten er språklig avvikende hos ham. Den beskriver et ungdomsmiljø, og språket preges av dette. *Vigdis Stokkelien* er likevel et moteksempel til denne sammenhengen, med en lav rangering på nr. 61 sammen med en avstand på bare 26 til forfatteren. Samtidig viser Figur 18 at Stokkelien er plassert i en tett normklynge, noe som gir mange konkurrenter. Dette reflekteres av den korte avstanden mellom vinnerne og Stokkeliens to tekster (7,4 og 7,6). *Olav Ottersen* har fire korrekte attribusjoner og ellers høye rangeringer mellom 2 og 4 ved de tekstene som ikke ble attribuert til ham, noe som indikerer relativt lite språklig variasjon (med hensyn til våre åtte språklige egenskaper) mellom hans tekster, sammenlignet med hva de øvrige forfatterne har. Alle hans tekster har forholdsvis liten avstand til forfatteren. Det kan ha oppveiet for at han er mer sentralt plassert i normklyngene enn de øvrige som har treff. Se avsnitt 4.4 for mer komplette data om dette. *Dag Solstad* har til dels høyere avstand mellom forfatteren og de tekstene der han er vinner, enn de øvrige forfatterne har, men han er også den som er plassert lengst unna normklyngen i fig. 18 (se også avsnitt 4.4), og har derfor færre konkurrenter. Han kommer ut som den mest sannsynlige forfatter av fem av sine seks tekster, men Bjørn Gunnar Olsen vinner ved teksten *T. Singer*. Vi så i fig. 12 (nr. 6) hvordan denne teksten har en avstand til Olsen på bare 1,0 i diagrammet over normklyngen i fig. 11, mens avstanden til Solstad var 8,4. I normklyngen i fig. 17 er det også *T. Singer* som er Solstad-teksten plassert like ved Olsen. Dette har sannsynligvis vært utslagsgivende.

*4.4 Faktorer som påvirker sjansen for korrekt forfatterattribusjon*
Vi har dermed identifisert tre faktorer med sannsynlig innflytelse på muligheten for korrekt forfatterattribusjon basert på våre åtte normrelaterte grammatiske valgmuligheter:

(a) forfatterens plassering perifert eller sentralt i en eller flere normklynger;

(b) avstanden mellom den undersøkte teksten og resten av forfatterens tekster i det flerdimensjonale rommet som defineres av de åtte grammatiske valgmulighetene – en avstand som avspeiler graden av språklig variasjon tekstene imellom;

(c) antallet tekster forfatteren er representert med i undersøkelsen, som kan påvirke faktor (b).

Så langt har vi dannet oss et inntrykk av faktor (a) for hver forfatter ved å betrakte forfatternes plassering i forhold til normklyngen i figur 18. Et mer komplett bilde av avstandene mellom forfattere må baseres på alle åtte egenskaper samlet. Resultatet av dette vises i tabell 3.

| Forfatter | 1<br>Snittavst. til 337 forf. | 2<br>Minsteavst. tekst/kompl. | 3<br>1 minus 2: attr.-sjanse | 4<br>Antall tekster | 5<br>% treff |
|---|---|---|---|---|---|
| **Solstad** | 83,2 | 10,2 | 73,0 | 6 | 83 |
| **Bye** | 68,5 | 8,4 | 60,1 | 6 | 67 |
| **Saus** | 68,5 | 17,4 | 51,1 | 4 | 75 |
| **Saabye Christensen** | 57,7 | 12,1 | 45,6 | 8 | 25 |
| **Ottersen** | 50,3 | 8,6 | 41,7 | 9 | 44 |
| **Jenseg** | 67,4 | 29,3 | 38,1 | 2 | 0 |
| **Sandberg** | 58,9 | 26,9 | 32,0 | 3 | 0 |
| **Lønnebotn** | 57,2 | 32,6 | 24,6 | 3 | 0 |
| **Stokkelien** | 50,5 | 26,0 | 24,5 | 2 | 0 |
| **Munck** | 59,9 | 38,0 | 21,9 | 2 | 0 |

*Tabell 3. Forfatternes sannsynlige sjanser for korrekt attribusjon (kolonne 3) basert på avstander mellom forfattere og minsteavstander mellom tekster i det åttedimensjonale rommet.*

I tabell 3 er forfatterne som hadde korrekte attribusjoner, markert med rødt. Kolonne 1 viser den gjennomsnittlige avstand mellom forfatteren og hver av de øvrige 337 forfatterne i det åttedimensjonale rommet der alle åtte egenskaper er med. Jo høyere denne avstanden er, desto mer pe-

rifert plassert er forfatteren i de ulike normklyngene, og desto større skulle sjansen være for korrekte attribusjoner. Kolonne 2 viser den minste avstanden mellom en tekst og resten av forfatterens tekster samlet (komplementmengden av tekster). Jo høyere denne avstanden er, desto mer varierer forfatteren sitt språk, og desto *mindre* skulle sjansen være for korrekte attribusjoner. Verdiene i kolonnene 1 og 2 trekker altså i motsatt retning av hverandre. I kolonne 3 er disse verdiene kombinert ved at minsteavstanden mellom tekstene er subtrahert fra avstanden mellom forfatterne, for å komme frem til en verdi som bedre skulle predikere sjansen for en eller flere korrekte attribusjoner. I tabellen er forfatterne ordnet etter verdiene i kolonne 3, som også er markert med rødt. Vi ser at Ottersen, som har den laveste avstand av samtlige til andre forfattere og derfor synes å være den gjennomsnittlig mest sentralt plasserte i normklynger, likevel havner så høyt som på femte plass i kolonne 3 fordi han også har en svært lav minsteavstand mellom tekstene sine, altså minst språklig variasjon mellom dem. Han har dessuten den klart laveste gjennomsnittlige avstanden mellom tekstene (ikke vist i tabell 3).

Alle fem forfattere med treff havner øverst i kolonne 3, og dermed også øverst i listen til venstre i tabellen. Snittavstand til andre forfattere, med subtraksjon av minsteavstanden mellom forfatterens tekster, synes altså å være en god indikator på sjansen til å oppnå korrekt forfatterattribusjon, med en grenseverdi omkring 40 i dette materialet.

Kolonne 4 antyder en sammenheng mellom antall tekster og minsteavstand mellom tekstene. Med flere tekster øker sjansene for at flere av dem ligger nær gjennomsnittet.

Kolonne 5 viser hvilken prosent av en forfatters tekster som er korrekt attribuert, og vi ser et visst samsvar med verdiene i kolonne 3.

Gjennomsnittsavstandene mellom hver av de 338 forfatterne og de 337 øvrige strekker seg fra 146,1 ned til 47,2. Solstad kommer med sin gjennomsnittsavstand 83,2 på 42. plass fra toppen. Ganske mange forfattere er med andre ord mer klyngeperifere enn han. Ottersen med 50,3 kommer på 323. plass.

## 5 Oppsummering og konklusjon

Denne studien har undersøkt tre morfologiske og fem syntaktiske valgmuligheter innenfor foreskreven og operativ norm i bokmål og fordel-

ingen av valgene blant 338 forfattere, basert på materiale fra trebanken NorGramBank. De åtte grammatiske valgmulighetene er listet opp først i avsnitt 2. Avsnitt 3 dokumenterer hvordan flere av disse valgene grupperer forfatterne og deres tekster i norm- og stilklynger, med varierende grad av korrelasjon mellom paret av egenskaper (der korr.=1 ville være perfekt positiv korrelasjon):

1. *enkel vs. dobbelt bestemthet og foran- vs. etterstilt possessiv: korr.=**0,71**;*
2. *-*en *vs. -*a *i mulige hunkjønnsord og enkel vs. dobbelt bestemthet: korr.=**0,17**;*
3. *-*en *vs. -*a *i mulige hunkjønnsord og foran- vs. etterstilt possessiv: korr.=**0,26**;*
4. en *vs.* ei *som ubestemt artikkel og -*et *vs. -*a *ved svake verb: korr.=**0,56**;*
5. være *vs.* ha *som hjelpeverb i perfektum av overgangsverb og enkel vs. dobbelt bestemthet: korr.=**0,39**.*

2, 3 og 5 har relativt lave korrelasjonsverdier, men grafene ved disse (fig. 10, 14 og 17) viser sammenhenger som ikke reflekteres tydelig i korrelasjonsverdiene. I grafene viser *x*-aksen prosentverdien for de syntaktiske trekkene foranstilt possessiv og enkel bestemthet. I alle tilfellene er det en tendens til at en høy verdi for disse trekkene impliserer en høy verdi for trekket vist langs *y*-aksen, mens det motsatte – at høy verdi for trekket langs *y*-aksen skulle tendere til å implisere en høy verdi for trekket langs *x*-aksen – ikke er tilfellet, og dette svekker korrelasjonsverdiene. Ved parene 2, 3 og 5 viser grafene altså at mye enkel bestemthet tenderer til å implisere mange *en*-endelser og mye bruk av *være* som perfektumshjelpeverb, og også at mye foranstilt possessiv tenderer til å implisere mange *en*-endelser, men ikke omvendt. Oppsummerende kunne man si at en formell, konservativ syntaktisk stil synes å favorisere en konservativ morfologi med *en*-endelser og trekket *være*-perfektum, mens en konservativ morfologi og *være*-perfektum på sin side ikke er bundet til en formell, konservativ stil, men brukes mer generelt.

Videre viser en kasusstudie av seks tekster av Dag Solstad, der vi tar for oss de grammatiske egenskapene enkeltvis eller parvis, hvordan få egenskaper av denne typen noen ganger kan identifisere en forfatter som forfatteren av en tekst blant flere hundre forfattere.

En undersøkelse av ytterligere ni forfatteres tekster, i tillegg til Solstads, der vi ser på alle åtte egenskapene samlet, tyder på at denne muligheten for korrekt forfatterattribusjon er påvirket av flere faktorer: forfatterens plassering perifert eller sentralt i en eller flere normklynger, avstanden mellom den undersøkte teksten og resten av forfatterens tekster i det flerdimensjonale rommet som defineres av de åtte grammatiske valgmulighetene – en avstand som avspeiler graden av språklig variasjon tekstene imellom – og antallet tekster forfatteren er representert med i undersøkelsen. Et mål på en forfatters plassering sentralt eller perifert i klynger er den gjennomsnittlige avstand mellom forfatteren og alle de øvrige forfatterne. Denne avstanden, med fratrekk av avstanden mellom forfatternes tekster, synes å være en god indikasjon på sjansen for korrekt forfatterattribusjon.

Hovedformålet med attribusjonsundersøkelsen har ikke vært å presentere en bedre metode for forfatterattribusjon enn de eksisterende, men å belyse den forfatter-individuerende evne som ligger i valgfriheten innenfor norsk skriftspråksnorm. Metoden presentert her vil neppe kunne konkurrere med statistikk over mer basale egenskaper, som for eksempel grunnleggende ordforråd og frekvenser av grammatiske formord, fordi den sistnevnte typen egenskaper er mer automatisert hos en skribent og derfor neppe varierer svært mye fra tekst til tekst hos samme forfatter. Norm- og stilegenskaper er sannsynligvis i større grad gjenstand for bevisste valg og kan derfor variere mer hos forfatteren, med usikrere attribusjon som resultat. Men det kan likevel være interessant å se nærmere på i hvilken grad slike egenskaper kan *supplere* de mer basale i attribusjonsforsøkene.

Undersøkelsen støtter antagelsen om at det rommet av valgmuligheter som åpnes av den spesielle norske skriftspråksituasjonen med stor valgfrihet, i noen grad fører til individuelle og gjenkjennbare normvalg som karakteriserer språket hos den enkelte forfatter. Samtidig viser normklyngene et mønster av mulige fremvoksende subvarieteter av bokmål som er av potensiell relevans for språklig rådgivning og fremtidig språknormering.

# 6 Appendiks

## *6.1 Forfattere*

Kriteriene for utvelgelse av undersøkelsens 338 forfattere er angitt til slutt i avsnitt 1.4. Oversettere av fremmedspråklige bøker er inkludert i listene, men ikke de fremmedspråklige forfatterne. Antallet kvinnelige og mannlige navn er ikke nøyaktig likt fordi trebankens 'forfatterenheter' ikke korresponderer én-til-én med forfatter- og oversetternavn. For eksempel kan flere oversettere ha oversatt samme forfatter, og samme oversetter kan ha oversatt flere forfattere.

### Kvinner:

| | | | |
|---|---|---|---|
| Aarø, Selma Lønning | Grorud, Cecilie | Lied, Solfrid Elgvin | Reitan, Karin |
| Aasen, Elisabeth | Grønli, Kristin Straumsheim | Lien, Margrethe | Riis, Annie |
| Abrahamsen, Aase Foss | Haff, Bergljot Hobæk | Lien, Merete | Ringen, Tone |
| Alfsen, Merete | Hagerup, Inger | Lindboe, Karin Kinge | Rygg, Pernille |
| Alver, Liv Margareth | Hansen, Mette | Lindell, Unni | Rypdal, Sylvi |
| André, Camilla | Hareide, Jorunn | Lindstrøm, Merethe | Røed, Tonje |
| Aspås, Anne-Berit | Haslund, Ebba | Lindvåg, Ellen Iris | Sandberg, Kristín A. |
| Baugstø, Line | Haugen, Eli | Lorck, Liv | Sanders, Hilde |
| Berentsen, Signy | Hauger, Torill Thorstad | Lorentzen, Karin | Scheen, Kjersti |
| Berg, Bjørg | Heide, Sigrid | Lund, May B. | Schjelderup, Daisy |
| Berg, Unni Marie | Heimvik, Abena | Lykken, Margaret Aronsen | Seaver, Kirsten A. |
| Berge, Mona | Helgesen, Helga | Lyngar, Mona | Siem, Inga |
| Berger, Marit | Henriksen, Vera | Løkkeberg, Vibeke | Skaar, Irene |
| Bergh, Kirsten | Hjorth, Vigdis | Lønnebotn, Ingri | Skjelbred, Margaret |
| Blomquist, Lise | Hoel, Kristin | Magerøy, Ragnhild | Smedstad, Tove Gravem |
| Bolin, Cathrine Bakke | Hoffengh, Sissel | Malling, Liv | Solås, Ragne |
| Bolstad, Kari | Hofsø, Ellen | Marstein, Trude | Sommerfelt, Aimée |
| Brekke, Toril | Hollup, Anne Grete | Mattsson, Annette | Spilde, Ingrid |
| Brenna, Gry | Holt, Anne | Mauno, Hanne | Steinslett, Kjellaug |
| Brodin, Elin | Ingulstad, Frid | McCrae, Kari | Stibolt, Helen |
| Bromark, Marit | Ingulstad, Tove | Moen, Grete Helene | Stokkelien, Vigdis |
| Bøge, Kari | Jakobsen, Hanne | Monsen, Nina Karin | Stubrud, Benedicta |
| Børresen, Beate | Jenseg, Grete Randsborg | Mulholland, Beate | Ståhl, Selma |
| Børsum, Lise | Jensen, Eva | Munck, Kirsten | Svendsen, Randi Berge |
| Christensen, Dorothea | Johansen, Janne Otnes | Mørck, Sidsel | Sverdrup, Kari |
| Christensen, Elsebeth | Johansen, Margaret | Nedreaas, Torborg | Sæther, Wera |
| Dahle, Gro | Karout, Lisbeth | Nerem, Marit | Sætvedt, Elisabeth |
| Dannevig, Tone | Kluge, Kirsti | Ness, Siri | Sævold, Ann Magritt |
| Eide, Elisabeth | Kristiansen, Nina | Nielsen, Unni | Sævold, Lisbeth |
| Eie, Ellen M. Haugen | Korssjøen, Kirsten | Nilsen, Ragnhild | Sævold, Magritt |
| Elligers, Anne | Krogedal, Else Lien | Nilsen, Tove | Sønsteng, Gry |
| Elstad, Anne Karin | Krohn, Anne-Berit H. | Nordahl, Marianne | Tandberg, Unni Wenche |
| Evensen, Iselin Røsjø | Køltzow, Liv | Nordby, Marit | Thorhus, Ann Mari |
| Fastvold, Marianne | Lange, Mona | Nortvedt, Reidun | Vestly, Anne-Cath. |
| Floer, Lisbeth | Lange-Nielsen, Sissel | Nævdal, Bodil | Vik, Bjørg |
| Fossum, Karin | Larsen, Britt Karin | Osmundsen, Mari | Waage, Hilde Henriksen |
| Frogner, Elsa | Larsen, Else | Oterholm, Anne | Wassmo, Herbjørg |
| Fugelsnes, Elin | Larsen, Kari E. | Pedersen, Bente | Winsnes, Else Marie |
| Fure, Paula | Larssen, Trude Brænne | Rabben, Vigdis | Wold, Kjersti |
| Gabrielsen, Berit Marianne | Lausund, Grete | Rafaelsen, Ellinor | Ørstavik, Hanne |
| Giske, Kari | Lerum, May Grethe | Ragde, Anne B. | |
| Grimsrud, Beate | Lie, Sissel | Raybo, Bjørg A. | |

**Menn:**

| | | | |
|---|---|---|---|
| Aarflot, Andreas | Espedal, Tomas | Kjærstad, Jan | Prytz, Carl Frederik |
| Aas, Erlend | Evensmo, Sigurd | Klippenvåg, Odd | Ramslie, Lars |
| Aavik, Asbjørn | Ewo, Jon | Knudsen, Sverre | Renberg, Tore |
| Alnæs, Karsten | Faldbakken, Knut | Knutsen, Per | Repstad, Pål |
| Ambjørnsen, Ingvar | Farbrot, Audun | Kolstad, Arild | Riisøen, Helge |
| Amdam, Per | Finne, Ferdinand | Kristensen, Vidar | Rimbereid, Øyvind |
| Andersen, Per Thomas | Fjørtoft, Kjell | Kristiansen, Idar | Risvik, Kjell |
| Angell, Olav | Flock, Willy | Kristiansen, Kristian | Rogde, Isak |
| Angell-Jacobsen, Rune | Fotland, Tor | Kvæstad, Jon | Rogstad, Anker |
| Askildsen, Kjell | Fretheim, Tor | Larsen, Terje Holtet | Rugstad, Christian |
| Aspeli, Widar | Gaarder, Jostein | Larssen, Vetle Lid | Rydningen, Tor |
| Bauer, Ola | Geelmuyden, Niels Chr. | Laukli, Svein-Arne | Rønning, Asle |
| Benestad, Finn | Graven, Andreas R. | Lie, Frank | Rønning, Bjarne |
| Berg, John | Groth, Henrik | Lie, Haakon | Rønning, Svend |
| Berggren, Arne | Gundersen, Gunnar Bull | Lillegaard, Leif B. | Røsholdt, Ole |
| Bing, Jon | Gunnerud, Jørgen | Loraas, Øivind | Saus, Anders |
| Brøgger, Waldemar | Haavardsholm, Espen | Lunde, Gunnar | Schjander, Nils |
| Hansen, Thore | Hagemann, Bror | Lønn, Øystein | Selmer, Odd |
| Bjerke, André | Hagerup, Helge | Madssen, Øivind | Skagen, Fredrik |
| Bjørklund, Ivar | Hagerup, Klaus | Martens, Johannes S. | Skogheim, Dag |
| Bjørneboe, Jens | Halstvedt, Tor | Mehlum, Jan | Skrede, Ingar |
| Bjørnstad, Ketil | Hamsun, Tore | Mehren, Stein | Solstad, Dag |
| Borgen, Johan | Haugen, Tormod | Michelet, Jon | Staalesen, Gunnar |
| Bottolvs, Bjørn | Haukås, Torfinn | Moe, Rolf Egil | Steen, Thorvald |
| Brandstadmoen, Geir | Havnes, Olaf | Molaug, Svein | Stenersen, Jan Erik |
| Brekstad, Kolbjørn | Havrevold, Finn | Myhren, Halvard | Stigen, Terje |
| Bringsværd, Tor Åge | Hoel, Sigurd | Mykle, Agnar | Svingen, Arne |
| Bugge, Niels Magnus | Holmås, Stig | Nilsen, Oddvar | Syvertsen, Håvard |
| Bye, Anders | Horvei, Nils | Nyquist, Arild | Sæterbakken, Stig |
| Bye, Skjalg | Hånes, Øivind | Nyrønning, Sverre M. | Søderlind, Didrik |
| Cappelen, Peder W. | Ingstad, Olav | Næss, Atle | Sørensen, Roar |
| Carling, Finn | Isaksen, Runo | Olsen, Bjørn Gunnar | Thommesen, Olav |
| Christensen, Arnfinn | Jacobsen, Roy | Olsen, Morten Harry | Torjussen, David |
| Christensen, Lars Saabye | Jakobsen, Ole Skau | Olsen, Pål Gerhard | Tufte, Leif |
| Dahl, Kjell Ola | Johansen, Knut | Omre, Arthur | Tunstad, Erik |
| Dahl, Tor Edvin | Johansen, Thorbjørn R. | Orvil, Ernst | Tusberg, Harald |
| Devold, Simon Flem | Johnsen, Victor | Ottersen, Olav | Tveit, Tore |
| Eidem, Odd | Johnsgaard, Magnar | Ottosen, Kristian | Ustad, Willy |
| Elster, Torolf | Jor, Finn | Pedersen, Einar | Vogt, Johan |
| Engelstad, Carl Fredrik | Jørgensen, Jan Christian | Pedersen, Erling | Vold, Torstein |
| Enger, Rolf | Jørgensen, Morten | Pedersen, Vidar | Winje, Trond |
| Eriksen, Erik | Kiøsterud, Erland | Poleszynski, Ernst W. | Wisløff, Carl Fr. |
| Eriksen, Trond Berg | Kjensli, Bjørnar | Rud, Nils Johan | Øgrim, Tron |

*6.2 Mulige hunkjønnsord*

De mulige hunkjønnsordene som er lagt til grunn for undersøkelsen av endelsen i bestemt form entall og av ubestemt artikkelform, er de som har mer enn 1000 treff i bestemt form entall i de relevante delene av trebanken (ikke begrenset til de 338 forfatterne). Det er følgende substantiver (der bare én variant nevnes ved hvert – 'fremtid' står for eksempel for både *framtid* og *fremtid*):

| | | | | | | |
|---|---|---|---|---|---|---|
| adresse | eske | handling | klokke | natt | setning | tro |
| avdeling | evne | havn | kone | nese | side | tunge |
| avis | ferd | helg | kulde | nærhet | sjel | tåke |
| befolkning | flaske | hensikt | kule | oppgave | skjorte | uke |
| behandling | fly | historie | kvinne | oppmerksomhet | skulder | ulykke |
| bestemor | forestilling | hjelp | kåpe | overflate | skyld | uro |
| bevissthet | forklaring | hud | lampe | pakke | slekt | utvikling |
| blokk | form | hule | leilighet | panne | slette | vakt |
| bok | forskning | hylle | linje | pike | sol | vekt |
| bro | fortelling | hytte | liste | pipe | sorg | venninne |
| brygge | fortid | hånd | lomme | plate | spenning | veske |
| bukse | fremtid | jakke | lue | presse | stemning | virkelighet |
| bygd | frihet | jakt | luft | pute | stillhet | vogn |
| bygning | frue | jente | lukt | regjering | stilling | åpning |
| celle | følge | jord | lykke | reise | strand | årsak |
| dame | gate | jul | løsning | rekke | stue | øy |
| datter | glede | kai | makt | retning | søster | |
| dronning | grav | kasse | mark | rolle | tante | |
| dukke | grense | kirke | maskin | rute | taushet | |
| dyne | gruppe | kiste | mening | sak | tekst | |
| dør | gul | kjærlighet | mor | sannhet | tid | |
| elv | hake | klasse | mulighet | seng | trapp | |

*6.3 Svake verb med den høyeste prosent* -a *i preteritum*

I avsnitt 2.3 nevnes grunnlaget for utvalget av de 196 svake verbene, av til sammen 1485 verb av denne klassen i trebanken, som har høyest prosent *-a* i hele korpus (ikke begrenset til de 338 forfatterne). Utvalget er gjort for å oppnå mer enn en forsvinnende andel *a*-endelser i undersøkelsen.

| | | | | | | |
|---|---|---|---|---|---|---|
| akke | esle | kakke | likne | rekne | sprade | terpe |
| aksle | fikle | kappe | lirke | ringe | sprake | tinge |
| amme | fikse | kike | lunte | rote | spraye | tippe |
| bable | filtre | kjefte | manne | ruse | sprette | tipse |
| bakke | fingre | kjekke | meisle | rusle | sprike | tisse |
| banne | fleipe | klabbe | messe | råke | spurte | trakke |
| baute | fløyte | klakke | mobbe | røske | stange | tralle |
| belje | forbanne | klinke | monne | røyke | steine | trene |
| berge | fordre | klubbe | more | score | steppe | trimme |
| bevæpne | forlove | knabbe | mulle | sikle | stime | trolle |
| blånekte | forpakte | knulle | måke | sjaue | stinke | tryne |
| bløffe | forstue | korke | nagle | skjene | vake | tråkle |
| bokse | funke | kose | narre | skrape | vakne | tufte |
| bolte | gaule | krangle | nistirre | skulke | vogge | tukle |
| bomme | geipe | kreke | nuppe | sladre | vralte | tukte |
| breie | gjære | kristne | pare | slafse | yppe | tulle |
| bråke | grille | kræsje | pigge | slokne | zoome | ture |
| båre | grise | kutte | pirke | sløve | streike | tusle |
| digge | hagle | kverke | pisse | smadre | streke | tørne |
| diske | haike | kverne | prate | smatre | stresse | tøve |
| disse | hauke | kvinke | prute | sneie | stulle | tøyse |
| drible | havne | kviskre | pælme | snerte | stunde | utmatte |
| drille | helle | kvitne | pønske | sniffe | sture | |
| droppe | hinke | kåre | ramle | snoke | surne | |
| dulte | hovne | kødde | rane | sone | sveipe | |
| dumme | innstifte | lappe | rape | sote | synde | |
| dundre | jobbe | leite | rappe | spikke | såpe | |
| dure | jukse | lene | rask | spikre | takle | |
| dysse | jumpe | letne | raute | spleise | tasle | |

*6.4 Overgangsverb med* være *eller* ha *som perfektumshjelpeverb*
I undersøkelsen av *være* vs. *ha* som perfektumhjelpeverb er det valgt ut 109 verb, opplistet nedenfor, der homonymi med *være* som passivhjelpeverb ikke påvirker resultatet. Dette forklares i avsnitt 2.7.

| | | | | | |
|---|---|---|---|---|---|
| ankomme | fordampe | hende | opprinne | skje | tilfryse |
| avgå | fordufte | hvelve | oppstå | skli | tilgro |
| avta | foregå | hvirvle | ose | skrumpe | tryne |
| begynne | forekomme | innkomme | lykkes | slukne | unnslippe |
| besvime | forfalle | innløpe | løye | sovne | utdø |
| blekne | forlise | inntre | minke | spakne | utebli |
| bli | forstumme | inntreffe | mykne | sprekke | utgå |
| blomstre | forsvinne | ise | mørkne | springe | utkomme |
| bortfalle | forulykke | kjølne | plane | stige | utløpe |
| dampe | fremkomme | klarne | rakne | stilne | vake |
| dovne | fyke | komme | ramle | stivne | vike |
| drysse | gjenoppstå | krepere | reise | strande | visne |
| dø | gli | krympe | resignere | strømme | vokse |
| eksplodere | glime | krype | revne | størkne | våkne |
| ende | gro | lamme | rinne | støve | |
| falle | gråne | lande | ruste | stå | |
| falme | gulne | lave | ryke | svinne | |
| flasse | gå | omkomme | råtne | svulme | |
| flyte | havarere | opphøre | sige | synke | |

## 7 Referanser

Bjerke, André. 1962. Samnorskfilologen. Publisert i André Bjerke: *Sproget som ikke vil dø*. Riksmålsforbundet, Oslo 1964.

Björnsson, Carl-Hugo. 1968. *Läsbarhet*. Stockholm: Liber.

Blatt, Ben. 2017. *Nabokov's Favorite Word is* Mauve. New York, London, Toronto, Sydney, New Delhi: Simon & Schuster.

Dyvik, Helge. 2003. Offisiell og ikke-offisiell språknormering – nyttig eller skadelig motsetning? I Helge Omdal og Rune Røsstad (red.): *Krefter og motkrefter i språknormeringa – Om språknormer i teori og praksis* s. 25–40. Kristiansand S.: Høyskoleforlaget AS.

Dyvik, Helge. 2012. Norm clusters in written Norwegian. I G. Andersen (red.) *Exploring newspaper language* s. 193–219. Amsterdam/New York: John Benjamins.

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse & Martha Thunes. 2016. NorGramBank: A 'Deep' Treebank for Norwegian. I Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (red.), *Proceedings of the Tenth International*

*Conference on Language Resources and Evaluation (LREC'16)*, s. 3555–3562, Portorož, Slovenia. ELRA.

Helset, Stig J. 2017. *Norm og røyndom. Ein statistisk studie av operative normer i det nynorske skriftspråket*. Ph.d.-avhandling, Universitetet i Bergen.

Helset, Stig J. 2018. Norm og røyndom – tilhøvet mellom fastsett og operativ norm i nynorsk. *Maal og Minne*, 110(1), s. 71–137.

Iversen, Ragnvald. 1944. Voksterlivet i Henrik Ibsens lyrikk. *Edda*, 1–2.

Kjetsaa, Geir. 1984. *The Authorship of the Quiet Don.* Oslo: Solum Forlag.

Lødrup, H. 2011. Hvor mange genus er det i Oslo-dialekten?. *Maal og Minne*, 103(2), s. 121–136.

Meurer, Paul. 2012. INESS-Search: A search system for LFG (and other) treebanks. I Miriam Butt and Tracy Holloway King (red.), *Proceedings of the LFG '12 Conference,* s. 404–421, Stanford, CA:CSLI Publications.

Mosteller, Frederick & David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA.

Omdal, Helge og Lars S. Vikør. 2002. *Språknormer i Norge. Normeringsproblematikk i bokmål og nynorsk.* Gjøvik: Cappelen Akademiske Forlag AS.

Rosén, Victoria. 2000. Er norsk et naturlig språk? I Øivin Andersen, Kjersti Fløttum & Torodd Kinn (red.): *Menneske, språk og felleskap*, s. 157–73. Novus forlag.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, & Helge Dyvik. 2012. An open infrastructure for advanced treebanking. I Jan Hajič, Koenraad De Smedt, Marko Tadić & António Branco (red.) *META-RESEARCH Workshop on Advanced Treebanking at LREC2012,* s. 22–29, Istanbul, Tyrkia, Mai 2012. European Language Resources Association (ELRA).

Rosén, Victoria, Helge Dyvik, Paul Meurer & Koenraad De Smedt. 2020. Creating and Exploring LFG Treebanks. I Miriam Butt & Ida Toivonen (red.) *Proceedings of the LFG'20 Conference*, s. 328–348. Stanford, CA: CSLI Publications. URL: http://web.stanford.edu/group/cslipublications/cslipublications/LFG/LFG-2020/lfg2020-rdmd.pdf

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3), s. 538–556.
Sundby, Nils Kristian. 1974. *Om normer*. Universitetsforlaget.
Troland, Victoria. 2015. *Hvem er forfatteren? – Stilometriske undersøkelser av norske prosatekster*. Masteroppgave i datalingvistikk og språkteknologi, Universitetet i Bergen.
Vannebo, Kjell Ivar. 1980. Om språkvitenskapens normbegrep. *Tijdschrift voor Skandinavistiek* 1, s. 3–23.
Vikør, Lars S. 2007. *Språkplanlegging. Prinsipp og praksis.* Oslo: Novus forlag.

**Abstract**

Norm clusters are groups of texts displaying shared choices among alternatives within a norm, forming 'clusters' within the space of possibilities because alternative combinations of choices are more rare. Based on material in Bokmål from the Norwegian treebank NorGramBank this study investigates the occurrence of norm clusters encompassing syntactic as well as morphological phenomena, and furthermore the possibility to identify the author of a text on the basis of the author's position relative to such clusters. In sections 2–3 correlations between eight grammatical phenomena among 338 authors are investigated in parallel with a case study of author attribution based on one author. In section 4 the attribution study is expanded through comparison with nine further, randomly selected authors.

The study documents correlations between morphological and syntactic phenomena. The case study demonstrates the possibility that an author in some cases may be identified uniquely among hundreds of others as the author of his or her texts on the basis of a few norm- and style-related properties. The study of nine further authors supports the assumption of a connection between this possibility and the author's position relative to norm clusters, but also indicates that other factors may be operative.

# Grammatiske fingeravtrykk

Helge Julius Jakhelln Dyvik
Universitetet i Bergen
Institutt for lingvistiske, litterære og estetiske studier
Postboks 7805
NO-5020 BERGEN
helge.dyvik@uib.no