

SMITTSOMME KORONAORD

KOENRAAD DE SMEDT

Universitetet i Bergen

SAMMENDRAG

Produktive sammensetninger bidrar betraktelig til danning av nyord i norsk. Denne studien undersøker bruken av nye sammensetninger med *corona-* eller *korona-* i nyhetstekster på nettet over en periode tidlig i 2020 da koronapandemien utviklet seg. Materiale fra Norsk aviskorpus ble brukt til en kvantitativ analyse på en tidslinje. Ordforrådet økte kontinuerlig og sterkt i flere måneder, og forholdet mellom ordtyper og ordforekomster antyder en økning i tematikkens bredde. I løpet av en halv måned endret stavemåten seg fra utelukkende *corona-* til overveiende *korona-*.

[1] INNLEDNING

Covid-19-pandemien, som ble erklært av Verdens helseorganisasjon 11. mars 2020, har allerede blitt gjenstand for mye forskning, ikke bare innen medisin, men også innen humaniora, psykologi, samfunnsvitenskap og etikk, blant annet. Det er blitt etablert spesialiserte tekstkorpus som belyser ulike typer diskurs, fra vitenskapelige artikler (Lu Wang et al. 2020) til tvitringer (Dimitrov et al. 2020). Slike korpus danner grunnlag for forskning og anvendelser, for eksempel innen terminologi og leksikografi (Paton 2020; OED Editorial 2020), men de aller fleste inneholder kun engelsk språk.

De siste tiårene er det blitt utviklet mange ulike korpus for norsk, ikke minst ved Tekstlaboratoriet ved UiO der Janne Bondi Johannessen har vært en stor drivkraft. Men hverken det banebrytende Oslo-korpus av taggede norske tekster (Johannessen et al. 2000) eller det nyere og mer balanserte Leksikografisk bokmålskorpus (Fjeld et al. 2020) som Tekstlaboratoriet gjør tilgjengelig gjennom CLARINO,¹ kan brukes til å forske på dagsaktuelle språkfenomener.

I motsetning til vanlige korpus, som har et stabilt utvalg fra en avgrenset periode, oppdateres *monitorkorpus* stadig og hyppig, noe som tillater monitorering av endringer i språkbruk nærmest i sanntid. *Norsk aviskorpus* er et stort monitorkorpus av norske nettaviser (Andersen & Hofland 2012). Dette korpuset er blitt brukt i flere studier av neologismer og importord (Andersen 2012). Blant disse er en undersøkelse av neologismer i form av kreative sammensetninger med forleddet *aske-* etter utbruddet av vulkanen Eyjafjallajökull på Island i 2010 (De Smedt 2012).

[1] <http://tekstlab.uio.no/clarino/>

Pandemien som spredte seg i raskt tempo i begynnelsen av 2020 ble en ny anledning til å granske aktuell språkbruk i Norsk aviskorpus. Viruset ble i første omgang omtalt som et nytt *coronavirus* og fikk senere det mer spesifikke navnet *2019-nCoV*, som etter kort tid ble endret til *SARS-CoV-2*. Det er likevel få som bruker dette spesifikke navnet på viruset. Sykdommen som i slutten av februar 2020 hadde nådd 43 land, inkludert Norge, fikk etter hvert navnet *covid-19* («coronavirus disease 2019»), men selv om dette navnet er blitt mer vanlig, kalles såvel sykdommen som viruset helt enkelt for *korona* på folkemunne.

Etter at sykdomsutbruddet ble erklært som en pandemi, antok konsekvensene for samfunnet raskt uante proporsjoner, slik at disse ble gjenstand for omfattende nyhetsspredning og diskusjon i mediene. Aviser, fjernsyn, radio og andre nyhetskanaler spilte en viktig rolle i formidling av myndighetenes omfattende tiltak for å hindre smitte. Snart ble påfallende mye språklig kreativitet observert, særlig i sammensetninger med *corona-/korona-* som forledd, for eksempel *koronapasienter*, *koronatilfellet* og *koronarammede*.

Denne studien ser nærmere på nye sammensetninger med forleddet *corona-/korona-* (både med og uten bindestrek) i artikler som er blitt publisert i norske nettaviser. Som vi skal se senere, har de aller fleste av disse sammensetningene aldri blitt brukt tidligere. Jeg har undersøkt i hvilket tempo slike nye sammensetninger dukket opp i publiserte nyhetstekster, hvor stor bredde det er i bruk av disse sammensetningene og hva slags variasjon det er i stavemåten (*c-* eller *k-*).

[2] DATA OG METODE

Bruk av ferske nyhetstekster med annotasjon av kilde og dato forutsetter i praksis enten direkte høsting fra nettaviser, eller tilgang til arkiver og korpus som kontinuerlig høster avisartikler og gjør disse søkbare gjennom et grensesnitt. Siden førstnevnte fremgangsmåte er ressurskrevende, ble bare aggregert materiale i arkiver og korpus vurdert brukt.

Norsk aviskorpus blir oppdatert hver natt gjennom automatisert høsting av tekster fra utvalgte norske nettaviser. Aviser som for tiden inngår i korpuset, er Adresseavisen (AA, Trondheim), Aftenposten (AP, Oslo), Bergens Tidende (BT, Bergen), Dagsavisen (DA, Oslo), Dagbladet (DB, Oslo), Dagens Næringsliv (DN, Oslo), Fædrelandsvennen (FV, Kristiansand), Nordlys (NL, Tromsø), Stavanger Aftenblad (SA, Stavanger) og Verdens Gang (VG, Oslo). Korpuset har en bokmålsdel og en nynorskdel. I denne omgangen valgte jeg å fokusere på bokmålsdelen, som er betydelig større.

Avisartiklene som blir høstet til korpuset blir automatisk rensset inntil tilnærmet ren brødtekst gjenstår, som sammen med publiseringsdato og kilde blir lagret. Fordi hver forekomst av ord har informasjon om dato og kilde, er det mulig å spore ord og uttrykk over tid og sammenligne ulike aviser. Dette gjør også at ethvert utvalg av materiale i prinsippet lett kan gjenfinnes, slik at forskning ba-

sert på korpuset skal være reproduserbar så lenge det er tilstrekkelig spesifisert hvordan materialet er selektert og bearbeidet.

Norsk aviskorpus er tilgjengelig i to nettbaserte korpusverktøy. Den ene versjonen er søkbar gjennom et grensesnitt basert på IMS Corpus Workbench (CWB) (Evert & Hardie 2011), som støtter ekstrahering av ord som samsvarer med et regulært uttrykk. Resultatene blir presentert i form av en konkordans i KWIC-format («keyword in context») på en nettside.² Det samme korpuset er også tilgjengelig i CLARINO gjennom Korpuskel,³ som har flere søke- og visningsmuligheter (Meurer 2012; se også Meurer, dette bind). Korpuskel-versjonen, som jeg valgte å bruke her, oppdateres ikke like often, men tillater enkel nedlasting av søkeresultater i form av en tabell med alle valgte attributter uten å måtte ekstrahere informasjonen fra en HTML-side med KWIC. Nedlasting fra Korpuskel kan gi et større antall treff enn det som kommer frem i CWB-versjonen, fordi sistnevnte fjerner konkordanslinjer som er duplikater.

Ved hjelp av søkekommandoen "[ck]orona.*" %c :: year = "2020" i bokmålsdelen fant jeg alle treff av ord som starter med *corona* eller *korona*, med store eller små bokstaver, publisert i året 2020. Selve ordet *corona/korona* og bøyingsformene til ordet ble filtrert ut, slik at bare sammensetninger gjenstod. Tastaturfeil og irrelevante ord som *kronar* og *coronal* ble fjernet ved hjelp av en manuelt laget liste. Den rensede listen har 75218 forekomster, som alle er sammensetninger. De første treffene var fra en artikkel med tittelen «Ny virusvariant kan stå bak mystisk sykdomsutbrudd i Kina» som ble publisert i VG 9. januar⁴ og innsamling av resultater var begrenset til 26. mai 2020. Det vil si at materialet inneholder treff fra en periode på 139 dager.

Dette datasettet ble filtrert og sortert på ulike måter gjennom en arbeidsflyt bestående av Unix shell-scripter og Awk-programmer. Resultatene ble videre aggregert og visualisert ved hjelp av det statistikkorienterte programmeringsspråket R. Av hensyn til automatisk sortering ble datoformatet *måned/dag* brukt i listene og grafene.

[3] KVANTITATIV ANALYSE

[3.1] *Stavemåte*

Latin *corona* («krans») har gitt opphav til flere norske ord og navn. Ordet *korona* med betydning «den synlige kransen rundt solen» er oppført med forbokstav *k*- i Bokmålsordboka, 2. utgave (Landrø & Wangensteen 1993). I denne utgaven finnes hverken oppslaget *coronavirus* eller *koronavirus*.

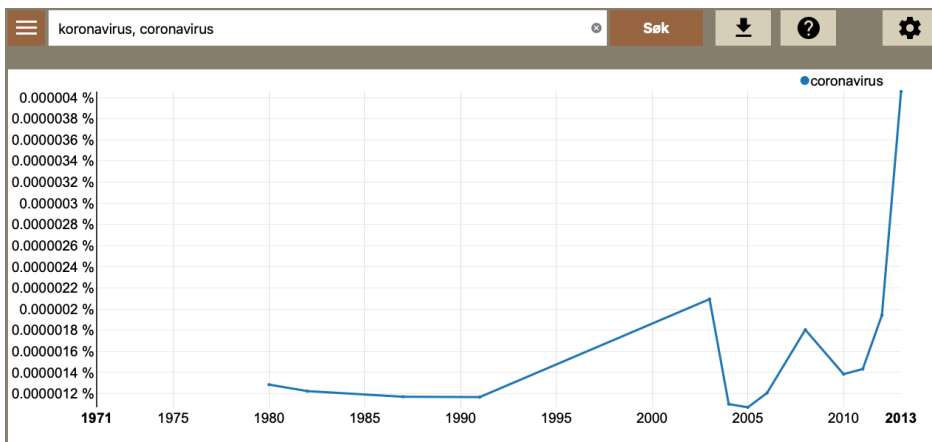
[2] <http://korpus.uib.no/avis/>

[3] <http://clarino.uib.no/korpuskel>

[4] Denne artikkelen ble tatt inn i Norsk aviskorpus på publiseringsdatoen, men versjonen på VGs nettsted ble oppdatert 24. januar 2020.

Norsk aviskorpus har før 2020 svært få sammensetninger med *corona*- som er relatert til viruset, nemlig *coronavirus(et)*, *coronavirus-familien*, *corona-familien*, *corona-smitte*, *corona-syk*, *coronavirussykdommen*, *corona-infeksjon* og *corona-vaksiner*.⁵ Alle disse ble skrevet med forbokstav *c*.⁶ Noen sammensetninger med *korona* forekommer før 2020 i andre betydninger. For eksempel ble *koronautbrudd* kun brukt om kraftige utladninger av plasma fra solens korona.

Også andre kilder brukte tilsynelatende utelukkende stavemåten *corona* før 2020. Et *n*-gramsøk i Nasjonalbibliotekets digitaliserte avissamling til og med 2013 viser at *coronavirus* bare forekommer med *c*-, både i bokmål og i nynorsk, som vist i figur 1. Stavemåten med *k*- fantes ikke.⁷ Tidsskrift for Den norske legeforening publiserte 21 artikler som nevner *coronavirus* i perioden 2003–2019, men ingen artikler med *koronavirus* før 2020.⁸



FIGUR 1: Relativfrekvenser i Nasjonalbibliotekets avissamling: *coronavirus* forekommer fra 1980 til 2013, *koronavirus* forekommer ikke.

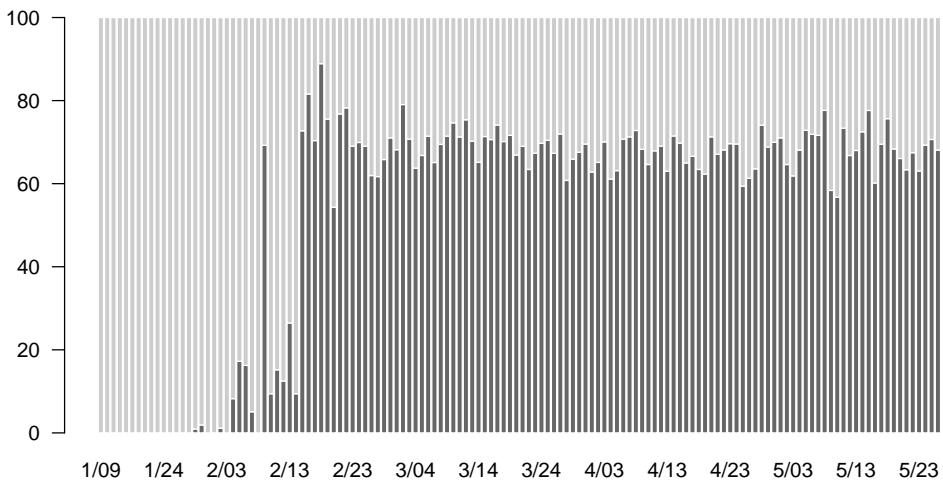
Den lange perioden med dokumentert uniform stavemåte med *c*- tok slutt i 2020 da Språkrådet gikk ut med et normerende utspill. I en artikkel om «Koronavirus og covid-19» skrev Språkrådet at ordet *korona* blir brukt både om viruset og sykdommen, og tilføyer «Vi skriv *korona* med *k*- på norsk, og samansetjinga *korona*-

- [5] Før 2020 dreide det seg om andre respiratoriske virus enn SARS-CoV-2, hovedsakelig SARS-CoV og MERS-CoV.
- [6] Ordet *korona-epidemien* forekommer i en artikkel publisert i AP 21. desember 2019, men denne forekomsten skyldes en oppdatering 30. mars 2020, se <https://www.aftenposten.no/sport/fotball/i/K32j9M/overgangsstudio-glimt-bekrefter-tuil-profilen-klar-for-klubben>, sist besøkt 10. september 2020.
- [7] <http://link.uib.no/korona-ngram>, besøkt 16. mai 2020.
- [8] <https://tidsskriftet.no/sok?&text=coronavirus> og <https://tidsskriftet.no/sok?&text=koronavirus>, besøkt 18. mai 2020.

virus er eit vanleg samnamn som skal ha liten forbokstav.»⁹ Språkrådet kontaktet NRK og Folkehelseinstituttet og informerte om den korrekte stavemåten.

Etter hvert ble oppslagsordet *koronavirus* lagt til NAOBs nettordbok, med *coronavirus* som variant.¹⁰ I denne ordboksartikkelen står både et sitat fra 2003 der ordet er skrevet med *c-* og et annet fra 6. mars 2020 med *k-*. Litt senere ble nettutgaven av Bokmålsordboka og Nynorskordboka oppdatert med en ny artikkel om *koronavirus* med *k-* i begge målformer.¹¹

Mot denne bakgrunnen har jeg kartlagt sammensetningenes stavemåte i det nevnte datagrunnlaget fra Norsk avis-korpus. Som figurene 2 og 3 viser, ble det funnet klar variasjon i stavemåte både over tid og mellom aviser.



FIGUR 2: Prosentvis fordeling mellom *c-* og *k-* over tid.

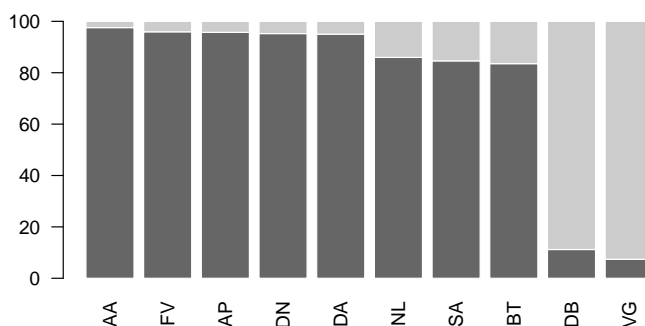
I januar ble nesten utelukkende stavemåten med *c-* brukt, men fra 29. januar ble enkelte tilfeller med *k-* observert; etter en kort periode med ustabilitet i første halvdel av februar kan vi merke oss noenlunde stabil bruk av *k-* i mellom ca. 60 og 70 prosent av tilfellene fra siste halvdel av februar. Denne endringen er visualisert i figur 2, mens figur 3 viser en klar todeling mellom aviser. Mens de fleste avisene i stor grad brukte stavemåten med *k-*, er det to aviser som skriver mest *c-*.

Disse funnene tyder på at det i løpet av februar har foregått en reell, dog ikke allment praktisert, endring i stavemåten. Til sammenligning viser resultatene fra et uformelt søk på nettstedet *nrk.no*, at *c-* var i bruk der i januar og så sent som 21.

[9] <https://www.sprakradet.no/Vi-og-vart/hva-skjer/Aktuelt-ord/koronavirus/>, publisert 28 januar, oppdatert 27. mars 2020, besøkt 29. mars 2020.

[10] <https://www.naob.no/ordbok/koronavirus>, besøkt 30. mars 2020; denne ordboksartikkelen ble publisert 16 mars ifølge leksikograf Carina Nilstun i NAOB (personlig kommunikasjon).

[11] <https://ordbok.uib.no/perl/ordbok.cgi?OPP=koronavirus>, besøkt 29. mars 2020; denne artikkelen ble publisert 23. mars ifølge Sturla Berg-Olsen i Språkrådet (personlig kommunikasjon).



FIGUR 3: Prosentvis fordeling mellom c- og k- per avis; prosent k- er angitt i mørk farge. AA=Adresseavisen, AP=Aftenposten, BT=Bergens Tidende, DA=Dagsavisen, DB=Dagbladet, DN=Dagens Næringsliv, FV=Fædrelandsvennen, NL=Nordlys, SA=Stavanger Aftenblad og VG=Verdens Gang.

februar,¹² men så tidlig som 28. januar ble også k- brukt, riktig nok i en artikkel der begge skrivemåter forekommer.¹³

[3.2] Frekvens og produktivitet

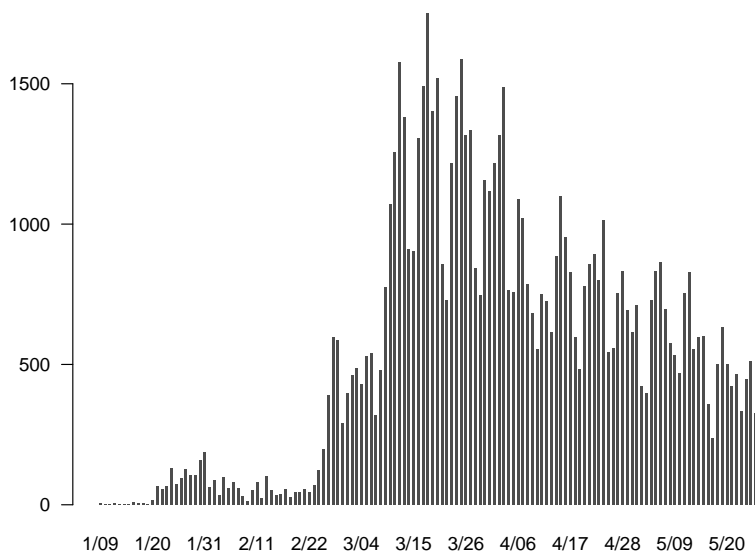
Antallet forekomster («tokens») per dag er vist i figur 4. De tidligste forekomstene av relevante sammensetninger i Norsk aviskorpus' 2020-del var *coronavirus* og *coronaviruset* i ovennevnte VG-artikkel av 9. januar 2020. Bruken av disse og andre sammensetningene i norske aviser forble beskjeden i en drøy måned, men antallet forekomster nådde over 200 per dag fra 26. februar, da viruset for første gang ble påvist i Norge. Rundt 12. mars, da strenge tiltak ble innført i Norge, kan vi merke oss en ny økning. Kort tid etter ble toppen nådd med 1752 forekomster på én dag. Tallene er noe lavere på lørdager og søndager siden færre artikler blir produsert i helgene.

Antallet forekomster sier noe om hvor mye det skrives om et tema, men for å granske bredden i tematikken må vi se nærmere på antallet ulike ord (typer). For å beregne typefrekvensene ble ordene normalisert ved at forleddet (*korona/corona*) og eventuell bindestrek ble fjernet;¹⁴ i tillegg ble store bokstaver konvertert til små. Deretter ble alle etterleddene lemmatisert, det vil si, bøyingsformer ble re-

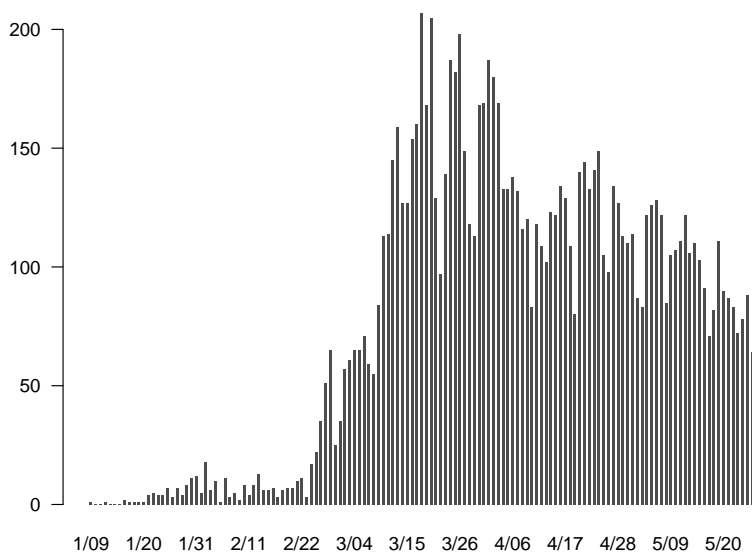
[12] <https://www.nrk.no/nordland/sykehusene-ma-spare-pa-verneutstyr-pa-grunn-av-corona-viruset-1.14913219>, publisert 28. januar 2020, oppdatert 4. mai 2020, besøkt 19. mai 2020.

[13] <https://www.nrk.no/ytring/slik-kan-coronaviruset-ramme-verdensokonomien-1.14879262>, publisert 6. februar 2020, besøkt 29. mars 2020.

[14] Blant totalt 75218 forekomster var det 9365 med bindestrek etter forleddet.



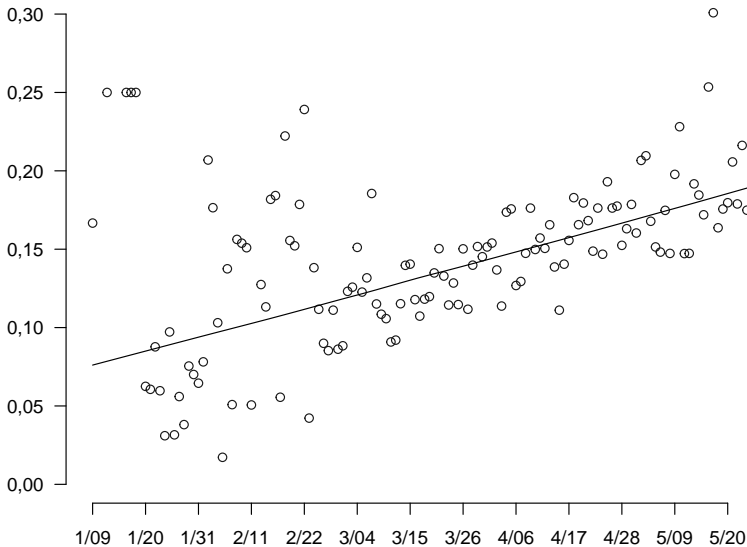
FIGUR 4: Antall forekomster observert per dato.



FIGUR 5: Antall typer (ulike lemna) observert per dato.

dusert til sin oppslagsform (lemma).¹⁵ Normalisering og lemmatisering gjør, for eksempel, at *koronatiltakene* og *Corona-tiltak* blir regnet som samme lemmatype *tiltak*, mens *oppmyking* og *oppmykning* forblir ulike typer. Lemmatisering er vanskelig, blant annet på grunn av tvetydigheter; *forneker* kan for eksempel være et substantiv eller en bøyingsform av *fornekte*. Noen små inkonsistenser påvirker likevel ikke det store bildet som tegnes i denne undersøkelsen.

De 75218 forekomstene fordeler seg over 1206 lemma. Jeg laget en frekvensliste¹⁶ som viser blant annet at lemmaet *-virus* som etterledd står for nesten halvparten av forekomstene; de nest frekvente ordene er *-krise* og *-smitte*. Det ikke-sammensatte ordet *corona/korona* og bøyingsformene, som ble holdt utenfor listen med sammensetninger, har en frekvens på 5617. Selve ordet er forresten en neologisme, i den forstand at *corona/korona* i denne betydningen ikke fantes tidligere utenfor sammensetninger som *coronavirus* og lignende.



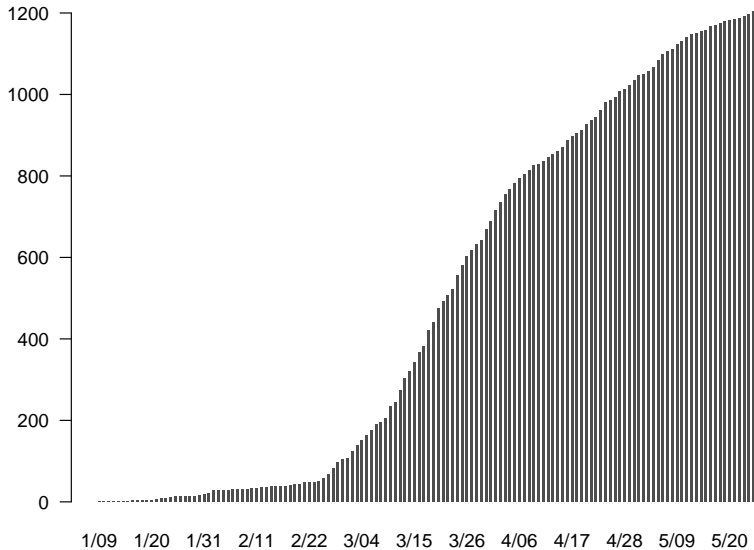
FIGUR 6: Variasjon (typer/forekomster).

Antallet typer (ulike lemma) per dag, som vist i figur 5, viser en fordeling over tid som ved første øyekast er i takt med antallet forekomster. For å se nærmere på dette forholdet mellom typer og forekomster, kan man i figur 6 se antall typer delt på antall forekomster. Trendlinjen viser en stadig økning av typer i forhold til forekomster, noe som tyder på at bredden i diskusjonen er blitt større. Tidlig

[15] Modellen *nb_core_news_md-2.3.0* i Python-modulen *spacy* ble brukt til dette formålet, se https://github.com/explosion/spacy-models/releases/tag/nb_core_news_md-2.3.0 og https://spacy.io/models/nb#nb_core_news_md. Resultatet av lemmatiseringen inneholdt noen feil og mangler som ble forsøkt dekkert mest mulig og korrigert manuelt.

[16] <https://github.com/clarino/corona>

i perioden svinger tallene ganske mye, men fra mars er det en mer sammenhengende økning som fortsetter mot slutten av perioden, selv om det fra april er en nedadgående tendens i antallet forekomster.



FIGUR 7: Kumulativ økning av ordforrådet.

Skal vi måle tempoet i økning av ordforrådet, bør vi se nærmere på antallet nye ord for hver dag, det vil si sammensetninger som ikke har forekommet på tidligere datoer i den undersøkte perioden. Den kumulative økningen av antallet nye ord er vist i figur 7. I januar og februar økte antallet sakte, mens en skarp akselerasjon kan observeres fra 26. februar da viruset nådde Norge. Fra dette tidspunktet har det vært en betydelig og konstant produksjon av nye sammensetninger, som begynner å flate litt ut i løpet av april.

I tillegg laget jeg en liste over alle ulike lemma som ble brukt som etterledd, sortert etter første dato og første avis de ble observert i.¹⁷ Denne listen viser altså i hvilken rekkefølge nye ord har kommet inn i ordforrådet.

[4] MORFOSYNTAKTISKE OG SEMANTISKE ASPEKTER

Den «opprinnelige» sammensetningen er *coronavirus*, men selv om denne er den mest frekvente, er det bare 99 av 1206 ulike typer i datamaterialet som inneholder *virus*. Forleddet *corona-/korona-* har tilsynelatende blitt frakoblet fra *-virus* og er blitt brukt til å referere ikke bare til viruset, men også til sykdommen, tiltak for å begrense smitte og de mange konsekvensene av disse tiltakene. Det fins noe

[17] <https://github.com/clarino/corona>

interessant i at *korona* er mer produktivt enn *virus*, for det jo vitterlig virusets virkning som er poenget, ikke at viruset ser ut som en krans. Men ordet *korona* er unikt, spesifikt og iøyefallende – noe som avisjournalistene ikke kvier seg for å bruke –, mens *virus* er mer generisk og nøytralt. *Koronahverdag* vinner altså på bekostning av *virushverdag* på bakgrunn av onomasiologisk saliens (jf. Geeraerts 1993). Fortolkningen av sammensetningene bygger ofte på en kjede av kausale og andre relasjoner. For eksempel kan ordet *koronahverdagen* beskrive vår hverdag som har endret seg som en konsekvens av restriksjoner for å bremse smitten fra viruset som tilhører koronavirusfamilien. Dette illustrerer at det er vanskelig med en streng komposisjonell tilnærming til sammensetninger og dermed understreker Saussures «tegnets arbitraritet».

De fleste sammensetningene har et substantiv som etterledd. Til sammen dekker disse et bredt spekter av betydninger. Mange sammensetninger er noenlunde transparente, men bare i kontekst av den kjente situasjonen. Før 2020 hadde det for eksempel vært vanskelig å tolke *koronatider*, *korona-17.mai*, *koronaperspektiv* og *koronautsettelsene*, mens *koronautbrudd* ble brukt i en annen betydning, som nevnt ovenfor. Blant de mer kreative og metaforiske nominale etterleddene er *-knekken*, *-knipen*, *-kluss*, *-smell*, *-spøkelse*, *-dilemma*, *-splitt*, *-psyken*, *-tabu*, *-koden*, *-grepet*, *-tsar*, *-sjokk*, *-tsunamien*, *-blemme*, *-tunnelen*, *-manøver* og (*kraftigere*) *-lut*.

Verbale etterledd er blant annet *-sikre*, *-stanse*, *-avlyse*, *-stenge* og *-teste*. Blant deverbale etterledd (partisipper brukt som adjektiv) kan vi merke oss *-smittet*, *-infisert*, *-merket*, *-trøblet*, *-relatert*, *-stengt*, *-rammet*, *-eksponert*, *-avlyst*, *-isolert*, *-stanset*, *-testet*, *-utsatt*, *-permittert*, *-kansellert*, *-bekymret*, *-tilpasset*, *-svindlet*, *-herjet*, *-påvirket* osv. Semantikken til disse etterleddene er relatert til hvordan smitte eller tiltak på grunn av smittefaren har påvirket mennesker eller virksomheter. Adjektiviske etterledd som *-fast*, *-fri*, *-syk*, *-frisk*, *-sårbar*, *-positiv*, *-ledig*, *-gavmild* og *-nervøs* viser ofte til tilstander hos mennesker.

I noen tilfeller virker sammensetningene overflødige. I en artikkel der pandemien er tema kunne man ha skrevet at «den bitre medisinen hjelper» istedenfor «den bitre koronamedisinen hjelper». Pandemien har ført til *restriksjoner* og *karantene*, men journalister nøler ikke med å skrive om *koronarestriksjoner* og *koronakarantene*, selv om konteksten er soleklar. Det virker som om produksjon og bruk av nye sammensetninger er blitt en smittsom tendens blant journalistene.

Det kan påpekes noe overlapp med ordene med forleddet *aske-* i 2010, diskutert i De Smedt (2012). Adjektivet *koronafast*, som ble observert allerede 28. januar 2020 og som forekom 227 ganger, ble sannsynligvis dannet etter *askefast*, som var relativt høyfrekvent i april og mai 2010 (114 forekomster) og som i sin tur hadde eldre sammensetninger, for eksempel *værfast* og *snøfast*, som modeller. Andre etterledd som kom frem i begge studiene, har en mer generell betydning som antyder et problem som sprer seg og som utfordrer oss, for eksempel *-krise*, *-kaos*, *-situasjon*, *-krisepakke*, *-spredning*, *-frykt*, *-rammet*, *-utbrudd* og *-fri*.

[5] DISKUSJON

Det er mye som tyder på at disse koronatider gir godt grunnlag for koronanyheter av dobbelt slag, både samfunnsmessige og leksikalske. Denne studien har sett nærmere på produktivitet med forleddet *corona-/korona-* i norske nettaviser i perioden fra 1. januar til 26. mai 2020. Sammensetningene med dette ordet utgjør et ganske spesielt fenomen i denne perioden, ikke bare fordi de forekom i ufattelige mengder, men også fordi de er så mange forskjellige og nesten alle er helt nye. Funnene viser ikke bare en høy og kontinuerlig produksjon av nye sammensetninger i en ganske lang periode, men kanskje det mest påfallende er at variasjonen øker stadig, noe som antyder at diskursen omfatter et stadig bredere spekter av situasjoner og hendelser.

Samtidig er det også mange nye kreative sammensetninger med mer vanlige forledd. Norsk aviskorpus har separate sider med nyord der man finner hundrevis av dem, for eksempel *hostebot, respirator Krise, en-meters-regelen, håndvaskopp-læring, hyttebestemmelsen, pandemimottak, testkø, isolasjonstider, respiratorkapasitet, massepermittering, flokkimmunitetsstrategien, publikumstomt, virus-utestengte, besøksstans, permitteringskrise, sanitærbarrierer, stengt-lappene, dopapirhamstring, fleggangsmunnbind, sy-dugnad* og *flyredningspakkene*. En nærmere granskning av slike nyord faller utenfor denne artikkelens rekkevidde, men det er kanskje verdt å nevne at *selvisolasjon* ser ut til å være en ny norsk sammensetning, mens engelsk *self-isolation* har blitt observert siden 1834 (Paton 2020).

Det store antallet sammensetninger med *corona-/korona-* i nyhetstekster kommer ikke helt uventet. Det kan være noe særlig journalistisk i å lage sammensetninger som ofte er redundante men som gir grunnordet en forsterkende effekt. Hvor bevisst valg det er av journalistene, vet vi ikke, men det kan bidra til å fange oppmerksomheten til leserne. Det er sikkert et poeng i å understreke at «også dette er en koronanyhet». Dette inntrykket stemmer overens med resultatene fra den tidligere studien av nye sammensetninger med *aske* som kom opp i samme korpus i 2010 (De Smedt 2012). Denne studien synliggjorde at journalister utviste stor kreativitet gjennom stadig nye sammensetninger over en 40-dagers periode med *aske* i atmosfæren.

Undersøkelsen beskrevet her har muligens noe samfunnsmessig relevans gjennom anvendelser, blant annet i leksikografiske og sosiologiske sammenhenger (jf. Paton 2020). Vi kan også se denne metoden i sammenheng med annen forskning, nærmere bestemt automatisk detektering av «burstiness», dvs. periodevis forhøyet forekomst av ord i en informasjonsstrøm (Kleinberg 2002). Kombinert med automatisk analyse av norske sammensetninger (Johannessen & Hauglin 1996), har slike metoder et potensiale for å automatisk fange opp hendelser fra norske nyhetstekster eller sosiale medier, også mindre hendelser som ellers ville gått under radaren.

I denne undersøkelsen har jeg tatt utgangspunkt i norske nettaviser som var

tilgjengelig i Norsk aviskorpus. Selv om dette utgjør et betydelig materiale fra ti viktige aviser på landsbasis, er mange andre aviser og skriftlige nyhetskanaler, blant annet nettstedet *nrk.no*, ikke representert. Norsk aviskorpus er likevel en gullgruve for undersøkelser av neologismer, ikke minst pga. at det i Korpuskel er godt tilrettelagt for korpuslingvistiske undersøkelser.

En annen mulig kilde var nyhetsarkivet Atekst Retriever,¹⁸ som innhenter materiale fra enda flere nyhetskilder enn Norsk aviskorpus. Det er likevel mer omstendelig å ekstrahere informasjon fra Retrievers grensesnitt siden det ikke gir strukturert informasjon i form av en konkordans eller tabell, men bare en rekke sammendrag av artikler. Dessuten oppdaget jeg umiddelbart noen problemer med datering. Figur 8 viser to utklipp av materialer fra Retriever. Det første klippet er datimerket 28. desember 2019.¹⁹ En nærmere kilde sjekk viser imidlertid at publiseringsdatoen i realiteten er 28. februar 2020.²⁰ Det andre klippet er tagget 5. desember 2019, men den arkiverte versjonen ble tilsynelatende oppdatert senere; kildens nettside ble sist oppdatert 10. april 2020.²¹



FIGUR 8: To skjermbilder med utklipp fra artikler arkivert i Atekst Retriever.

Eksempler på «stavemåterevisjonisme» ble funnet på *nrk.no*, for eksempel i en artikkel der *corona-* forekom så sent som 22. april 2020, men ble endret til *korona-*

[18] <https://web.retriever-info.com/services/archive>, besøkt 17. mars 2020.

[19] Denne artikkelen ble senere tilsynelatende fjernet fra arkivet.

[20] <https://www.lp.no/nyheter/sykehuset-klart-for-koronapasienter/>, også publisert som <https://www.agderposten.no/nyheter/sykehuset-klart-for-koronapasienter/>, besøkt 17. mars 2020.

[21] <https://blikk.no/film-film-og-tv-lesbisk/charlies-lesbiske-engel/183185>, besøkt 19. mai 2020.

noen dager senere.²² Slike forhold skaper krøll når man vil plote historiske forekomster på en tidslinje, som er målet i denne studien. Norsk aviskorpus innhenter, tagger og lagrer tekstene senest dagen etter publikasjon. I korpuset blir datoen for hver artikkel forsøkt ekstrahert fra opptil tre spesifikke steder i materialet (Andersen & Hofland 2012, s. 13). Denne metoden fungerer bra så lenge informasjonen i kildematerialet kan tolkes utvetydig. Mer generelt er den høyst reelle muligheten for at feil sniker seg inn i aggregerende arkiver og korpus, noe som korpuslingvister, medievitere og andre bør være oppmerksom på. Denne faren er også blitt påpekt i forbindelse med digitaliserte bøker (Nunberg 2009).

Denne undersøkelsen har sine begrensninger når det gjelder tekstkilder. Hitil mangler en analyse av tilsvarende sammensetninger i nynorskdelen av Aviskorpuset. Det hadde også vært interessant å se på andre kommunikasjonskanaler, for eksempel radio, fjernsyn og sosiale medier. Likeså hadde det vært interessant å foreta en sammenligning med andre språk som har produktive sammensetninger. Et kjapt blick på nederlandske og tyske medier avslører at det også på disse språkene er stor leksikalsk koronakreativitet.

CLARIN, den europeiske forskningsinfrastrukturen for språkressurser og -teknologi, har laget en oversikt over tilgjengelige aviskorpus.²³ De fleste av disse er dessverre ikke monitorkorpus, men er begrenset til en avsluttet periode. Dessuten er disse ressursene ikke harmonisert for koding og annotasjon, slik at det gjenstår mye arbeid for å fremme kompatibilitet. Det er rom for mange flere koronastudier, også på tvers av sjanger og språk, som faller utenfor denne artikkelens rekkevidde.

TAKKSIGELSER OG MINNEORD

Takk til Knut Hofland for teknisk utvikling og oppdatering av Norsk aviskorpus, og til Paul Meurer for tilgjengeliggjøring gjennom Korpuskel i CLARINO. Takk til Mikkil Ekeland Paulsen, Carina Nilstun, Margunn Rauset, Victoria Rosén, Sturla Berg-Olsen og redaksjonskomiteen for kommentarer.

Janne Bondi Johannessens bortgang føles som et stort tap. Hun var en viktig bygger av norsk språkinfrastruktur. Jeg savner hennes smittsomme latter.

REFERANSER

Andersen, Gisle (red.). 2012. *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian* (Studies in Corpus Linguistics 49). Amsterdam: John Benjamins.

[22] https://www.nrk.no/anmeldelser/anmeldelse_-_jeg-foeslar-at-vi-vakner_av-beate-grimsrud-1.14991444 publisert 22. april 2020, oppdatert 27 april 2020.

[23] <https://www.clarin.eu/resource-families/newspaper-corpora>

- Andersen, Gisle & Knut Hofland. 2012. Building a large corpus based on newspapers from the web. I Gisle Andersen (red.), 1–28.
- De Smedt, Koenraad. 2012. Ash compound frenzy: A case study in the Norwegian Newspaper Corpus. I Gisle Andersen (red.), 241–255.
- Dimitrov, Dimitar, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zsu, Matthäus Zloch & Stefan Dietze. 2020. TweetsCOV19 – A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic [Preprint]. I *29th ACM International Conference on Information & Knowledge Management (CIKM2020), Resource Track*, Association for Computing Machinery. doi:10.1145/3340531.3412765. <https://arxiv.org/pdf/2006.14492.pdf>.
- Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. I *Proceedings of the Corpus Linguistics 2011 conference*, Birmingham, UK. <https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf>.
- Fjeld, Ruth Vatvedt, Anders Nøklestad & Kristin Hagen. 2020. Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk. *OSLa* 11(1). 101–124. doi:10.5617/osla.8176.
- Geeraerts, Dirk. 1993. Generalised Onomasiological Salience. *Belgian Journal of Linguistics* 8(1). 43–56. doi:10.1075/bjl.8.04gee. <https://www.jbe-platform.com/content/journals/10.1075/bjl.8.04gee>.
- Johannessen, Janne Bondi & Helge Hauglin. 1996. An Automatic Analysis of Norwegian Compounds. I *Papers from the 16th Scandinavian Conference of Linguistics*, redigert av Timo Haukioja, 209–220. Turku/Åbo, Finland. <https://www.duo.uio.no/handle/10852/43896>.
- Johannessen, Janne Bondi, Anders Nøklestad & Kristin Hagen. 2000. A web-based advanced and user friendly system: The Oslo corpus of tagged Norwegian texts. I *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, <http://www.lrec-conf.org/proceedings/lrec2000/pdf/363.pdf>.
- Kleinberg, Jon. 2002. Bursty and Hierarchical Structure in Streams. I *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 91–101. doi:10.1145/775047.775061.
- Landrø, Marit Ingebjørg & Boye Wangensteen (red.). 1993. *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Oslo: Universitetsforlaget.

- Lu Wang, Lucy, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni & Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset [preprint]. *ArXiv* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/>.
- Meurer, Paul. 2012. Corpuscle – a new corpus management platform for annotated corpora. I Gisle Andersen (red.), 31–49.
- Nunberg, Geoffrey. 2009. Google’s Book Search: A Disaster for Scholars. *The Chronicle of Higher Education* <https://www.chronicle.com/article/Googles-Book-Search-A/48245/>.
- OED Editorial. 2020. Corpus analysis of the language of Covid-19. <https://public.oed.com/blog/corpus-analysis-of-the-language-of-covid-19/>.
- Paton, Bernadette. 2020. Social change and linguistic change: The language of Covid-19. <https://public.oed.com/blog/the-language-of-covid-19/>.

SUMMARY

Productive compounding contributes a lot to neology in Norwegian. This study investigates the use of new compounds with *corona-* or *korona-* occurring in online news articles in early 2020 when the coronavirus pandemic developed. Material from the Norwegian Newspaper Corpus was used in a quantitative analysis on a timeline. The vocabulary found to increase continually and strongly during several months, and the proportion of word types to tokens indicates a broadening of themes. After barely half a month the spelling changed from only *corona-* to mostly *korona-*.

KONTAKT

Koenraad De Smedt
Universitetet i Bergen
desmedt@uib.no